NMR METABOLIC PROFILING OF MOSQUITO SPECIES TO UNDERSTAND INSECTICIDE RESISTANCE

Thesis submitted in accordance with the requirements of the

University of Liverpool

for the degree of

Doctor in Philosophy

by

RUDI GROSMAN

MAY 2019

I. Acknowledgments

First and foremost, my family hence this following paragraph will be in Turkish.

Her şeyden önce annem Liza ve babam İzak, her ne kadar bana her şeyi kendi başıma başardığımı tekrarlasanızda, siz olmadan bugüne kadar başardığım hiçbir şeyin mümkün olamayacağını ve aksini iddia etmenin tamamen bir deli saçması olduğunu bilmenizi isterim. Sizi çok seviyorum ve sonsuz kez teşekkür ediyorum. Yosef, yıllarca bizimkiler ve benim armada jenerasyon tercümanlığı yapmasan, yediğim birçok haltın altından bu kadar kolay kalkabileceğimi hiç sanmıyorum ve balkonda viski akşamlarını çok arıyorum. Her zaman yanımda olup destek çıktığın için çok teşekkür ederim. Özellikle isimle teşekkür etmek istediğim son kişi dayım Aron'cukum. Çekirdek ailemden sonra en büyük destekçimsin benim geleceğime dair hedeflerin benimkilerden büyük olduğu aşikâr, desteğin için cok teşekkür ederim. Son olarak burada adi geçmeyip bana destek olan eski ve yeni aile fertlerine sonsuz teşekkür ediyorum, iyi ki varsınız, iyi ki yanımdasınız.

I would like to extend my gratitude to some key people in IIB and LSTM who made this project possible. Firstly, my supervisors Lu-Yun Lian and Andy Jones for giving me the opportunity to undertake this project. Helping and supporting me throughout this project. As well as all the interesting and helpful discussion along the way. Many thanks to Gareth Lycett for allowing me to use the insectary in LSTM and many helpful discussions for the thesis to materialise. Amalia Anthousi for showing me the ropes of mosquito breeding and bearing with me while I took pictures. Special thanks to Marie Phelan for challenging every idea I proposed in NMR metabolomics and Igor Barsukov for all the interesting conversations on NMR, interesting approaches and applications. Eva Caamaño Gutierrez and Arturas Grauslys for their immense help in stats as well as many nights of food, drink and board games! Everyone in Lab-C and NMRC past and present; fellow students, sleepless post-docs and tireless technicians.

Finally, I would like to thank few people who helped me to stay sane and entertained from the day I met them. Stephanie Yip, Chris Janot and Manuel Barday thank you all for being great friends on good and bad days; always remember the year of 1884. Lastly, a very big thank you to Kathryn Williams. There were many bad days consoled by you, many setbacks you helped me through and many sleepless nights you accompanied me. Thank you for all the support that you have given me.

II. Abstract

The work presented in this thesis explores insecticide resistance using NMR metabolomics in pupa, early pupa and early adult mosquitoes of both sexes. Firstly, sex differences were investigated in early pupa and early adult strains of *Anopheles gambiae* and *Aedes aegypti*. Secondly, Cyp4g16 and Cyp4g17 knock-down strains of *An. gambiae* were studied to further understand the cuticular hydrocarbon (CHC) biosynthesis mechanism. Lastly, wild type susceptible and resistant strains of *An. gambiae* and *Ae. aegypti* were studied to understand metabolic differences in resistance. A common theme in this work is the utilisation of NMR metabolomics, focusing on polar metabolic profiles and its application in insecticide resistant mosquitoes. Improvements on the analyses of the NMR data to make the identification of metabolites from the NMR spectra more robust was also undertaken in the course of this project.

The vast majority of mosquito studies focus on females since only females bite. This means that there is often a great amount of information on male mosquitoes that is not utilised. In order to investigate sex differences in wild type *An. gambiae* (and knock-down strains) and *Ae. aegypti*, a metabolomics protocol was established. This study found that early female pupae and adults have higher levels of lactate and glucose compared to males, whereas, their pyruvate and propionate levels were found to be lower.

Cuticular resistance is a relatively newly discovered mechanism, with much about it that is still unknown, Electron microscopy studies have shown that insecticide resistant mosquitoes of *An. gambiae* possess a thicker cuticular layer. Furthermore, the decarbonylases, Cyp4g16 and Cyp4g17, have been found to be critical in CHC biosynthesis and are highly expressed in insecticide resistant strains. Although not fully characterised, these enzymes are thought to catalyse a decarbonylation step in CHC biosynthesis. The study here showed evidence that both Cyp4g16 and Cyp4g17 take part in the biosynthesis of methyl branched-alkanes. Additionally, these enzymes have temporal activity where Cyp4g16 activity can be observed in both pupa and adult stages, whereas Cyp4g17 activity is only observable in the adult stage.

Using resistant and susceptible strains of both *An. gambiae* and *Ae. aegypti*, metabolic differences were observed. Trehalose was found consistently higher in all resistant strains

and so it has been proposed as a biomarker candidate. Furthermore, systems using trehalose were suggested as potential targets in insecticide development.

In conclusion, the work in this thesis shows that it is possible to use NMR metabolomics to study the metabolic responses found in insect pupae and adults, to distinguish sex differences and the obtained development stage- dependent temporal information on the activity of specific enzymes. Furthermore, through the identification of elevation in trehalose in resistant mosquitoes, this study has proposed a potential biomarker for identification of resistant mosquito species.

III. Abbreviations

ABC	ATP-binding cassette
AMIX	Analysis of mixtures
ANOVA	Analysis of variance
APCI	Atmospheric pressure chemical ionisation
APPI	Atmospheric pressure photoionisation
ARSyN	ASCA removal of systemic noise
ASCA	ANOVA simultaneous component analysis
BH	Benjamini-Hochberg
BMRB	Biological magnetic resonance bank
CDC	Centers for diseases control and prevention
CHC	Cuticular hydrocarbon
CHIKV	Chikungunya virus
CNS	Central nervous system
CoA	Coenzyme A
CONT	Control
CPMG	Carr-Purcell-Meiboom-Gill
CRS	Correlation reliability score
CYP	Cytochrome P450
DA	Discriminant analysis
DAVID	The Database for Annotation, Visualization and Integrated Discovery
DDT	Dichlorodiphenyltrichloroethane
DENV	Dengue virus
DSS	4,4-dimethyl-4-silapentane-1-sulfonic acid
EASE	Expression analysis systemic explorer
ECL	Even chain length
EDR	Embryonic desiccation resistance
EI	Electron ionisation
ESI	Electrospray ionisation
FAB	Fast atom bombardment
FDR	False discovery rate
FID	Free induction decay
GABA	gamma-Aminobutyric acid
GC	Gas chromatography
HC	Hydrocarbon
HHFW	Half height full width
HILIC	Hydrophilic interaction chromatography
HMDB	Human metabolome database
HPLC	High pressure/precision liquid chromatography
HSD	Honest significant difference
HSQC	Heteronuclear single quantum coherence

IQR	Interquartile range
ITN	insecticide-treated nets
JH	Juvenile hormone synthesis pathway
KD	Knock-down
KEGG	Kyoto encyclopaedia of genes and genome
LC	Liquid chromatography
LED	Longitudinal encode decode
LSTM	Liverpool school of tropical medicine
MALDI	Matrix assisted-laser desorption/ionisation
MDA	Mass drug administration
MS	Mass spectrometry
MSEA	Metabolite set enrichment analysis
MSI	Metabolomics standards initiative
MVAP	Mevalonate pathway
NA	Not applicable
NADPH	Reduced nicotinamide adenine dinucleotide phosphate
NIRD	Near infrared spectroscopy
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser effect
NOESY	Nuclear Overhauser spectroscopy
NS	Not significant
NST	No specific treatment
OCL	Odd chain length
PC	Principal component
PCA	Principal component analysis
PDB	Protein data bank
PLS	Partial least square
PLS-DA	Partial least square discriminant analysis
PNS	Peripheral nervous system
POR	NADPH-cytochrome P450s oxidoreductase
PPE	Personal protective equipment
PQN	Probabilistic quotient normalisation
PVCA	Principal variance component analysis
QA	Quality assurance
QC	Quality control
QIT	Quadrupole ion trap
QTOF	Quadrupole time of flight
QUAD	IN THE MS TABLE IN CHAPTER 1
RF	Radio frequency
RNA	Ribonucleic acid
RO	Reverse osmosis
ROC	Receiver operating characteristic
RVF	Rift valley fever

RVFV	Rift valley fever virus
SNR	Signal to noise ratio
SVA	Surrogate variable analysis
TOCSY	Total correlation spectroscopy
TOF	Time of flight
TSP	3-(Trimethylsilyl)propionic-2,2,3,3-d4 acid sodium salt
TSW	Tonic salt water
UAS	Upstream activation sequence
UPLC	Ultra-high pressure/precision liquid chromatography
USD	United states dollar
VIP	Variable importance of the projection
VK	Valle Kisumu
VLC	Very long chain
WHO	World health organisation
WT	Wild type
ZIKV	Zika virus

IV. Table of contents

Chapter 1

1 Introduction

1.1 Mosquitoes and mosquito borne diseases

Mosquitoes (*Culicidae* family) are one of the families described in the Diptera order [1]. The *Culicidae* family comprises approximately 3,500 two-winged species of mosquitoes. Almost all aquatic habitats in the world provide a breeding site for mosquitoes [2], except permanently frozen areas. Mosquitoes can inhabit humid tropics, warm moist climates, temperate and cool zones [3]. High variation in mosquito species adaptations enable them to inhabit temporary, permanent, highly polluted or clean water bodies [2].

Mosquitoes are mostly known for their vector role in disease transmission. As a vector, certain mosquitoes can host certain pathogens and take part in the pathogen's life cycle. These pathogens can cause a variety of diseases such as chikungunya, dengue fever, lymphatic filariasis, zika, and malaria. Generally, when a mosquito hosts a pathogen, the mosquito is not affected by it, but rather transmits the pathogen to a secondary host, such as humans, that will suffer the associated disease.

Although mosquitoes are not the only vectors for disease-causing pathogens, they cause the most damage to public health in terms of deaths caused [4]. Mosquitoes also account for the majority of the most closely monitored diseases. Figure 1.1-1 shows the percentage of diseases transmitted by insect vectors according to the World Health Organisation (WHO).

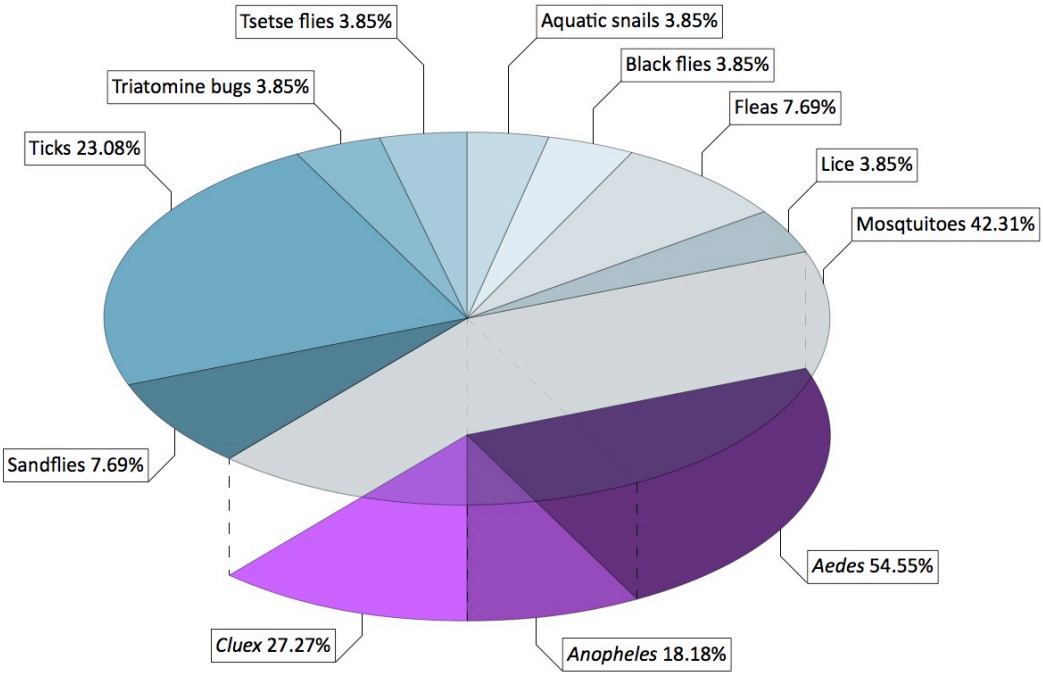


Figure 1.1-1: WHO most closely monitored diseases' vector distribution. Out of the vector borne diseases, mosquitoes are responsible for transmitting 42.31%. Data from WHO global vector control response 2017 – 2030 [4].

Within the mosquito family, different mosquito species host different pathogens and hence cause different diseases (Table 1.1-1). Among the most notable are malaria and zika.

Table 1.1-1: Most closely monitored mosquito vectors and their associated diseases [4].

Host genus	**Pathogen type**	**Pathogen name**	**Disease**
Aedes	Parasite	*Filarioidea spp*.	Lymphatic filariasis
	Virus	Chikungunya virus	Chikungunya
		Dengue virus	Dengue fever
		Phlebovirus	Rift Valley fever
		Yellow fever virus	Yellow fever
		Zika virus	Zika
Anopheles	Parasite	*Filarioidea spp.*	Lymphatic filariasis
		Plasmodium spp.	Malaria
Culex	Parasite	*Filarioidea spp.*	Lymphatic filariasis
	Virus	Japanese encephalitis virus	Japanese encephalitis
		West Nile virus	West Nile fever

Cumulatively, mosquito-borne diseases shown in Table 1.1-1 account for approximately one million deaths per year. On an economical perspective, the global direct cost of malaria alone is estimated around 12 billion USD per year.

An. gambiae and *Ae. aegypti* species were identified as the most influential due to the fact that *Anopheles* species account for the most deaths and *Aedes* species account for the greatest number of diseases contracted.

1.1.1 Impact on population

In developed countries, mosquitoes do not pose an endemic danger as much as they do in less developed countries. Approximately 17% of all infectious diseases are caused by vector borne diseases, causing 700,000 deaths per annum [4]. In 2016, malaria alone was estimated to cause 445,000 deaths [5]. Figure 1.1-2 shows countries in Africa, south Asia and south America with at least one malaria incident. Diseases with lesser deaths per year such as Chagas disease [6], leishmaniasis [7] and schistosomiasis [8] affect hundreds of millions of people worldwide. While there are numerous diseases of concern transmitted by *Ae. Aegypti*, none of them cause as many deaths as malaria. Figure 1.1-3 shows the current state of reported *Ae. aegypti*-borne diseases.

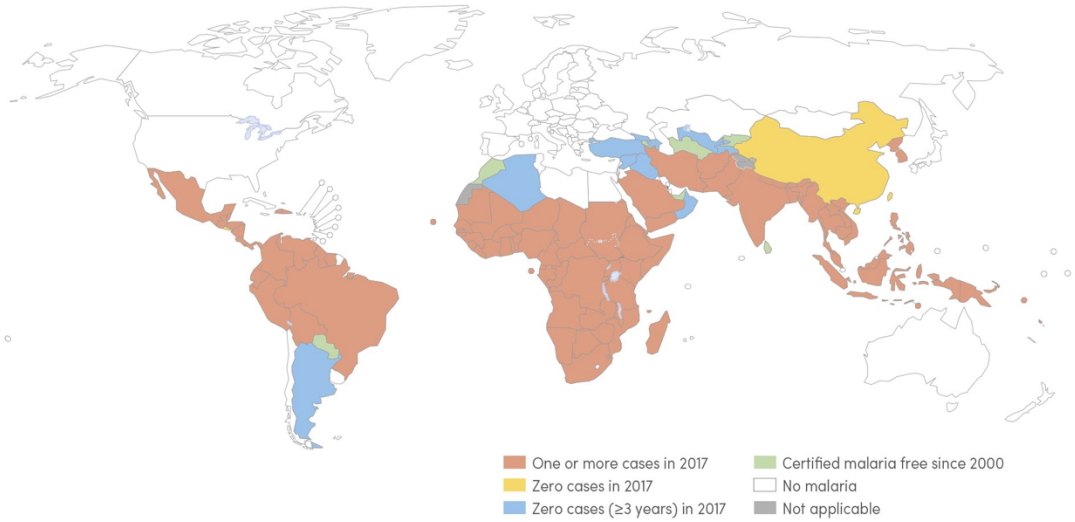


Figure 1.1-2: Malaria status changes between 2000 and 2017 according to reported cases. From the map it can be easily seen that countries around the southern hemisphere are more prone to have reports of malaria [5].

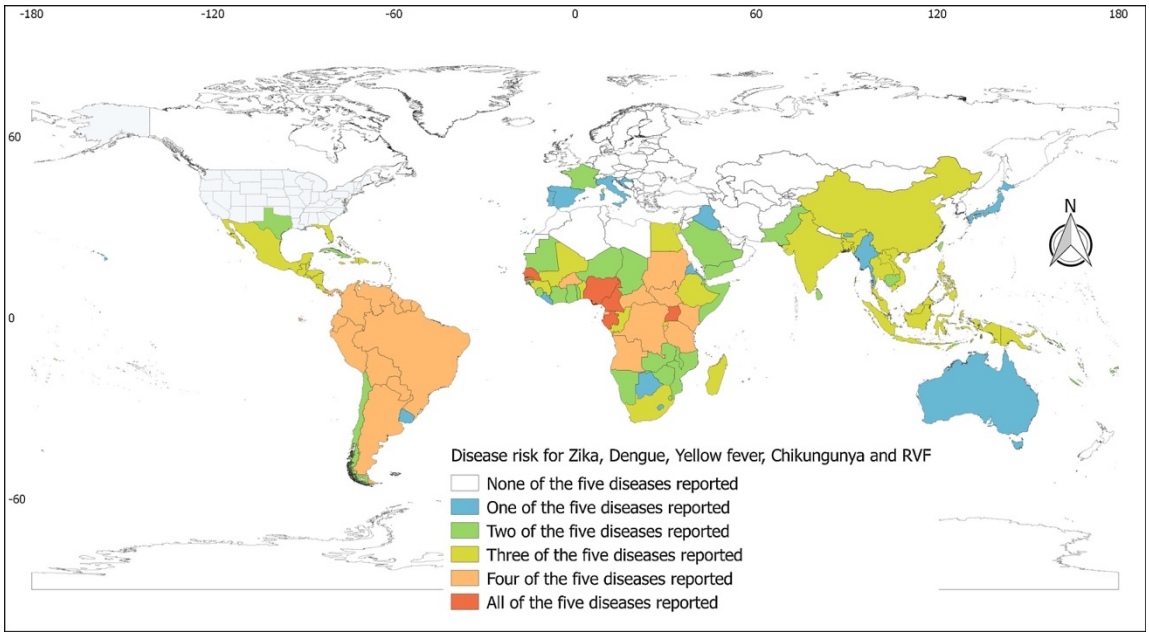


Figure 1.1-3: Map of *Ae. aegypti* disease reports for zika, dengue, yellow fever, chikungunya and RVF (Rift Valley Fever). White represents no reported cases and grey highlights the states of USA with no reported cases [9].

The burden of mosquito-borne diseases are not only measured by the deaths disabilities, and infections caused, but also by the serious financial impact they have on countries. For example, one study estimated that in Asia it would cost approximately 1,151 USD for treatment and related costs to treat a child infected with Japanese encephalitis [10]. To put this amount into context, the families in the study worked in rural areas in parts of Asia and had incomes of about 115 USD per month [10]. Other mosquito-borne diseases incur similar costs. The economic impact of West Nile virus in the United States between 1999-2012 was approximately 56 million USD per year [11], excluding vector control programs or money lost as a consequence of workforce sickness. In 2005, a West Nile virus outbreak in California was estimated to cost 2.98 million USD [12]; comparatively the 2002 outbreak in Louisiana, USA had a total cost of 20.14 million USD [13]. Furthermore, studies conducted on the cost of Dengue Fever from eight countries in the Americas and Asia (Brazil, El Salvador, Guatemala, Panama, Venezuela, Cambodia, Malaysia and Thailand) and Americas overall, estimates overall direct and indirect costs between 1.8 and 2.1 billion USD each year [14] [15]. Considering that there are 100 million infections per year, this disease has a high impact on economy and health care. In the case of Rift Valley Fever virus, an additional impact to consider is that it affects both humans and animals. Currently, it is most prevalent in Africa. The outbreak in Kenya between 2006 and 2007 resulted in a financial impact of 9.3 million in USD due to the effect on both people and livestock [16]. Moreover, to date, Yellow Fever is the only disease in Table 1.1-1 for which a commercial vaccine is currently available. Between 2011 and 2015, 330 million USD was spent on yellow fever vaccines in endemic countries

[17]. Other outbreaks such as the Chikungunya outbreak of 2017 in Bangladesh created a minimum loss of 606 USD per month per patient, caused by loss of productivity [18]. Finally, non-lethal malaria treatment alone can be a heavy burden on public health funds. In 2013, the cost of uncomplicated malaria treatments in Nigeria was 182,953.65 USD from a single hospital [19]. This sum accounted for 25% of the hospital's total expenditure of 2013 [19].

1.1.2 Diseases, treatment and prevention

Disease treatments most frequently treat the symptoms rather than the disease itself. As shown in Table 1.1-2 the majority of these diseases cannot be readily treated directly and do not have available vaccines for prevention. When a disease without a treatment is contracted, pathogen clearance is reliant completely on the host immune system. Meanwhile, only the symptoms presented, such as headaches, swelling and fever, can be treated. This approach is effective only in so much as the resources available in the region and the education level of the population in danger. Furthermore, immunosuppressed and vulnerable members of society such as those who are pregnant, juveniles and the elderly will have greater difficulties in clearing the pathogen.

Table 1.1-2: List of diseases monitored by WHO. MDA, mass drug administration; NST, no specific treatment; Anb, antibiotics; VC, vector control; Vac, vaccination; M, million; T, thousand; PPE, personal protective equipment.

Disease	Pathogen Type	Year	Cases	Deaths	Treatment	Prevention
Lymphatic filariasis	Parasite	2014	120 M	-	MDA	MDA ,PPE & VC
Chikungunya	Virus	2006	1.4 M	-	NST	PPE & VC
Dengue fever	Virus	Per year	100 M	12,500	NST	PPE & VC
Rift Valley fever	Virus			-	NST	PPE & VC
Yellow fever	Virus	Per year	200 T	30,000	NST	Vac, PPE & VC
Zika	Virus			-	NST	PPE & VC
Malaria	Parasite	2012	207 M	627,000	Anb	PPE & VC

To reduce the number of cases and the costs incurred by treatments, disease prevention methods play a critical role. Currently there are two widely adopted prevention methods: application of insecticides and use of insecticide-treated nets (ITN). Although the latter application mostly affects female mosquitoes, both rely on use of effective insecticides. Treatments for some of the more significant mosquito-borne diseases are described briefly below.

1.1.2.1 Lymphatic filariasis treatment

Lymphatic filariasis (also known as elephantiasis) is caused by parasites of *Filarioidea* species and can be transmitted by female mosquitoes of the species *Anopheles*, *Aedes*, *Culex* or *Mansonia* [20]. Lymphatic filariasis is known to be endemic in over 70 countries however

70% of the infected cases are in India, Nigeria, Bangladesh and Indonesia. Male to female ratio of infected patients is 10:1 which may be due to women's more covered attire in the highly endemic regions. Amongst the infected patients about 50% are between 30-50 years old.

Lymphatic filariasis can be diagnosed through testing of blood for microfilariae presence. A skin biopsy may be required if the parasite is a skin dwelling type. Symptoms of lymphatic filariasis can be divided into three groups: asymptomatic, chronic and acute [20]. Approximately half of the lymphatic filariasis patients are clinically asymptomatic. This is despite having the parasite in their blood. Depending on patient's immune response length of asymptomatic period can vary. Acute and chronic filariasis manifests itself by lymph gland inflammation episodes and swelling of the limbs and usually accompanied by pulmonary eosinophilia, fever and malaise [20]. Typically, fevers caused by lymphatic filariasis is accompanied with headache and chills.

The parasites causing Lymphatic filariasis can be eliminated from a patient's system with proper treatment with antibiotics and anthelmintics. Treatment for lymphatic filariasis is administered in the form of prophylactic chemotherapy to the endemic population [21]. Patients suffering from muscle deformation caused by the disease may experience disability and reduced life quality [20]. While these patients can undergo surgery to improve their life quality, the muscle deformation will not be fully cured and will require lifelong care. There are no commercially available vaccines for lymphatic filariasis, although a vaccine reported in 2013 was found to be effective in a *Rhesus macaque* model although a human trial has not been reported yet [22].

1.1.2.2 Chikungunya treatment

Chikungunya is a viral disease transmitted by *Ae. aegypti* females. Chikungunya Virus (CHIKV) is an *Alphavirus* [23]. Chikungunya was first described in 1952 in Tanzania and shortly after was identified in central and southern Africa as well [23]. Two outbreaks in Asia was recorded in 1958 and 1973 [23]. Following the 2004 outbreak in Kenya the virus spread to India and southeast Asia [23].

Chikungunya is diagnosed through serological tests and virus isolation for identification. The majority of Chikungunya patients recover fully but some may have symptoms persisting over

a longer time [23]. Chikungunya symptoms are fever, rash and severe aches in joints that usually progresses to a chronic stage [23]. More severe cases can cause extreme pain in hands, fingers and elbows hence debilitating patients [23]. Joint symptoms are typically characteristic to chronic form of the diseases [23]. Chikungunya is rarely fatal where it is mostly observed in elderly and new born. Approximately 50% of acute chikungunya patients have cutaneous manifestations in palms, soles of the feet, torso and face [23].

There is no specific treatment for Chikungunya [23]. In case of an infection, the symptoms of the disease are treated until it is cleared [23]. Although a commercial vaccine is not available there is a vaccine (MV-CHIKV) in phase II of clinical trials, started in August 2016 and expected to be completed by June 2018 [24]. In November 2018, Phase-II clinical trial of MV-CHIK was reported to be successful and Phase-III trials are currently being designed [25].

1.1.2.3 Dengue fever treatment

Dengue virus (DENV) is a *Falvivirus* and *Ae. aegypti* females are the primary vector for it, although, it can be transmitted by other *Aedes* species as well [26]. DENV transmission can be found in Eastern Mediterranean, American, South east Asia Western Pacific and African regions [26]. Highest DENC infection is in Asian children between ages 5 and 15 followed by American tropics between the ages of 19-40 [26].

Dengue fever can be diagnosed through serological test and DENV can be detected in the blood prior to 1-2 days before the fever onset and the following 5-6 days [26]. Different phases of the disease can be detected through immunoglobin M (acute) and G (recovery) detection [26]. Typically, dengue fever lasts a week, although in this time frame the disease can rapidly worsen. Dengue fever can be categorised in three phases: acute febrile phase, critical phase and recovery phase. Symptoms manifest in the acute phase as fever, headache, body pain and vomiting [26]. Most patients enter the recovery phase within 5 days and the symptoms deteriorate [26]. Some patients, worse following the fever drop and enter the critical phase [26]. Patients in this phase can experience shock induced by plasma leakage. Symptoms of this shock includes coldness, weak pulse, tachycardia and hypotension[26]. Dengue fever is another disease where no specific treatment exists. All treatment methods are focused on relieving the symptoms.

About 3.9 billion people are at risk of Dengue infection [27]. The first Dengue vaccine was licenced in 2015 by Sanfoni Pasteur [28], [29]. The vaccine was deployed in highly endemic areas on conditional recommendation by 2016 [30]. In 2017, further analyses showed the vaccine to be highly effective only in people who had previously contracted the virus [30]. Moreover, and in contrast, if the vaccine was used prior to any DENV infection, the patient actually had a higher risk of severe Dengue fever upon DENV infection through a mosquito [30].

1.1.2.4 Rift Valley fever treatment

Rift Valley fever virus (RVFV) is a *Phelebovirus* and transmitted primarily by *Aedes* females [31]. RVFV was first isolated in Kenya 1930 and is prevalent in Africa [32]. Recorded major outbreaks of RVFV were South Africa in 1950-19501, Egypt in 1977-1979, Mauritania in 1978 and Madagascar in 1990 [31]. In year 2000 first outbreak outside of Africa was recorded in Arabian Peninsula [31]. Humans as a host for RVFV is considered to be dead, due to low amplification of virus in human hosts [31]. Hence, completing the life cycle of the virus through a human host is favourable. Nevertheless, humans can be infected through mosquitoes and cattle [31].

Similar to other diseases RVFV can be detected through serological test by means of Immunoglobin M and G detection although this is usually possible between 4 to 7 days after infection which is also the typical duration of the diseases. Rift Valley Fever can be observed in two forms mild and severe. Mild form being the most common, manifests symptoms of fever, muscle & joint pain and headache [31]. The severe form of the disease typically manifests 1 or more of the following 3 syndromes: eye disease (0.5-2% of patients), inflammation of cerebral tissue (<1% of patients) and haemorrhagic fever (<1% of patients). For mild forms of Rift Valley fever, no treatment is required [31]. In severe cases the treatment is in the form of supportive therapy where symptoms are treated while the immune system clears out the viral infection [31]. Rift Valley fever can also infect livestock, which can have a high economic impact [32].

A vaccine for humans does not exist, although, there are vaccines available for livestock. Since cattle is the primary host that complete the host-vector cycle, immunising cattle plays an important part in reducing the number mosquitoes with RVFV hence reducing the number

of infections in humans. Hence there is great interest in developing better vaccines for livestock which also help with reducing the number of infected humans [32].

1.1.2.5 Yellow fever treatment & prevention

Yellow fever virus is a *Flavivirus* and transmitted by *Aedes* and *Hemagogus* females [33]. Yellow fever endemic is endemic in South America (most notably Brazil) and central Africa [33]. The "Yellow" in Yellow fever refers to the jaundice caused by the liver damage [33].

Yellow fever diagnosis can be carried out by serological test by detection of Immunoglobiln M [33]. Yellow fever is an acute illness with symptoms; fever, nausea, vomiting, pain in torso [33]. Yellow fever can cause hepatitis with jaundice, renal failure, haemorrhage, shock and can lead to death [33]. There is no direct treatment for Yellow fever hence, only the symptoms are treated. It is the only disease in Table 1.1-2 for which an effective vaccine is available. A single dose of the vaccine can immunize a person for a lifetime [33].

1.1.2.6 Zika treatment

Zika is caused by the *Flavivirus* zika virus (ZIKV) and is transmitted primarily by *Ae. aegypti* females [34]. Since it's firs known outbreak in 2007 in island of Yap, in the western Pacific, Zika virus was only recorded as epidemics [34]. Between 2007-2017 outbreaks were recorded in French Polynesia, south pacific & pacific islands, Brazil and Americas [34]. A survey conducted on the serological tests in French Polynesia showed 80% percent of the infections to be asymptomatic of which 30% were infants and 50% were adults [34].

ZIKV is diagnosed through serological tests where specific immunoglobin M antigens are used as a marker for the virus [34]. Due to immunoglobin M antigens not always being detectable complementary methods might be required for diagnosis and need to be carried out within the first 12 months of the infection [34]. Combination of urine analysis was shown to increase the detection rate ZIKV within the first week after onset of the symptoms [34]. ZIKV disease symptoms can appear between 3-12 days and are fever, joint pain, red eyes, headache and rashes [35]. There are no specific treatments available for Zika. During pregnancy, the Zika virus can be transmitted to the foetus which may cause birth defects such as microcephaly [36]. Following the relatively recent outbreak in Brazil, a reduction in Zika infections were observed. This is most likely due to the high media coverage the virus took in 2016, prior to the Olympics held in Brazil. As a consequence of the high media exposure, general public was

more aware of and took precautions for Zika virus [37]. Currently, there are no commercially available vaccines for Zika. A few are in clinical trials, although the decrease in number of Zika infections may provide a challenge in completing the trials [38].

1.1.2.7 Malaria treatment

Malaria is caused by the parasites of the *Plasmodium* genus [39]. Malaria endemic regions can be found in Africa, Central & South America and South-East Asia. In 2016, Malaria caused an estimate of 445,000 deaths[5]. Within this genus severe human malaria is mainly caused by *P. falciparum* [40]. *Anopheles* females are the vectors for the disease.

Gold standard for malaria diagnosis is investigating blood samples with light microscopy. Malaria disease can be categorised as uncomplicated or severe. This disease is mostly known for its severe cases which can be lethal; however uncomplicated malaria is treatable [41]. Uncomplicated malaria can be easily treated with anti-malarial drugs and manifests symptoms of fever, headaches, nausea and body aches [42]. The disease is mostly known for it severe malaria which can be lethal [42]. Severe malaria occurs when patient develops organ failures, blood and/or metabolic abnormalities [42]. Complications that can observed in a severe malaria patient are: impairment of consciousness, seizures, coma, anaemia, respiratory distress, low blood pressure, high acidity in the blood and low blood glucose [42].

Antimalarial drugs and treating the symptoms are the current way to deal with malarial infections [41]. However, resistance to antimalarial drugs is increasing [43] and so there is significant interest in preventing infection in the first instance. Currently there is only one vaccine that has shown preventative results, RTS,S/AS01 [44]–[46]. In clinical trials RTS,S/AS01 demonstrated an efficacy of approximately 20-45% when given four doses (up to 6 in babies) [44]. Although there is a potential vaccine in trials there is a challenge with the vaccine's receptivity. It has been reported that community engagement was inadequate due to lack of information about the vaccine [45]. Moreover some studies highlighted the fear of vaccine's side effects [45]. Additionally, inefficient delivery to children and low quality health service [45].

1.2 Mosquitoes *An. gambiae* and *Ae. aegypti*

Mosquito species of the *Culicidae* family can be found in fresh or salt-water marshes, mangrove swamps, rice fields, grassy ditches, the edges of streams as well as in small

temporary water collections [1], [2]. Even though some species may prefer certain breeding sites over the others, overall most water bodies can become a breeding site for mosquitoes. The life cycle of the different mosquito species is well conserved. Although there are differences between species, all mosquitoes go through a life cycle that comprises three aquatic (egg, larva & pupa) and one terrestrial (adult) stage (Figure 1.2-1).

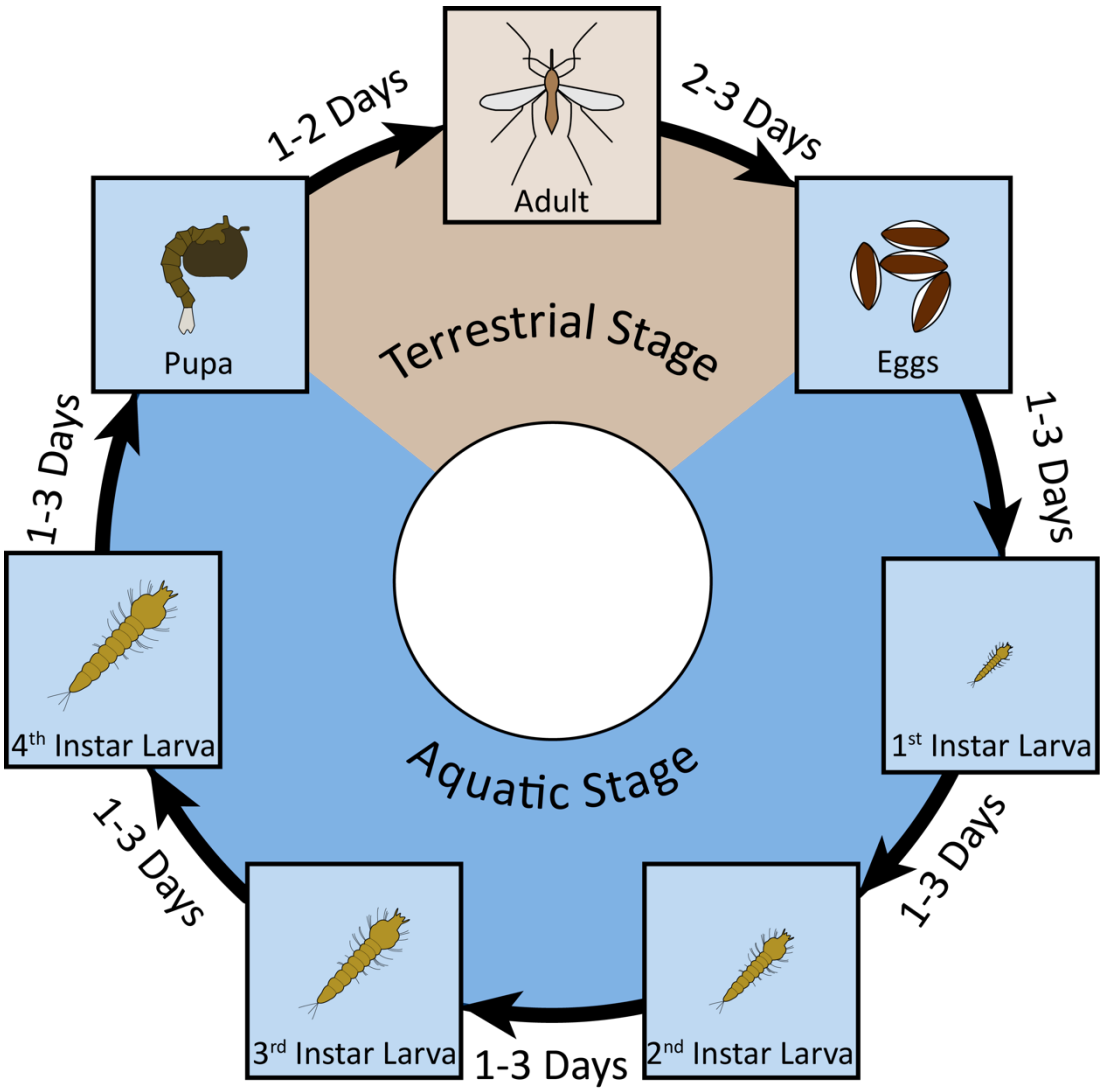


Figure 1.2-1: Mosquito life cycle through four main stages. The first three stages are in an aquatic environment (egg, larvae, pupa) followed by an adult stage in the terrestrial environment. Four instar stages make up the larvae stage progression. Eggs are typically laid on water bodies, although some species' eggs undergo quiescence which allows them to lay eggs on moist soil. Typically, eggs hatch within three days. The larval stage is a mobile stage, where larva swim and look for food. Through a series of moulting, larvae go through four sub-stages called instars. Each instar takes roughly three days. After the fourth instar, a pupa develops from each larva. The pupal stage is the last aquatic stage, where the pupae are still mobile in water. In this stage they do not feed, but stay close to the surface to breathe. Typically, in three days, adult mosquitoes emerge from the pupal casing. After a brief period of drying, they start to fly. Adult mosquitoes spend the rest of their life in a terrestrial environment. Upon successful mating, females return to a suitable breeding site to lay eggs, thus completing the life cycle.

Female mosquitoes lay between 50-200 eggs per oviposition [2]. Generally, mosquito eggs can be divided in to two groups depending on whether they enter a quiescence or not [2].

Egg quiescence is also known as embryonic desiccation resistance (EDR) and depends on a variety of factors such as eggshell structure & composition and metabolic activity of the larvae reside within the egg [47]. There are 3 layers to the egg shell critical for EDR: serosal cuticle (inner layer), endochorion (middle layer) and exochorion (outer layer) [47]. The serosal cuticle secretes a coating material under the chorion layer containing chitin which protects the embryo from desiccation. Additionally, changes in the abundance of the eggshell components such as hydrocarbons affects the water loss regulation [47]. Higher chitin content in eggshells increase desiccation resistance [47].

Species where eggs do not go through quiescence are laid directly on water [48]. Depending on the species, these eggs may be laid singly, in small groups (Figure 1.2-2), or in a raft form [49]. These eggs start their development immediately after oviposition, and hatch 2-7 days later depending on temperature. Species laying eggs who enter quiescence do not lay directly on water [2]. These eggs are laid singly on moist soil, which floods when water levels rise. During the quiescence, the eggs are protected against desiccation that may be caused by the arid conditions or elevated temperature [47].

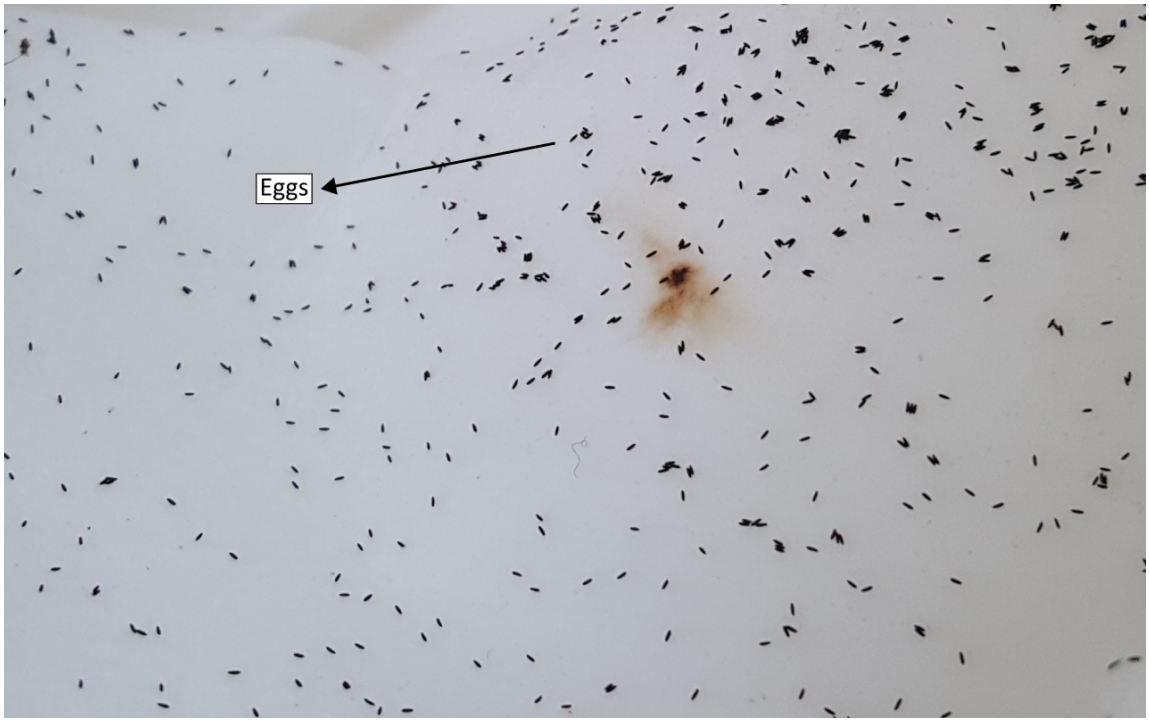


Figure 1.2-2: *Ae aegypti* eggs are laid singly or in small groups. These can undergo quiescence, which makes them more resistant to desiccation compared to *Anopheles* eggs. During quiescence, the eggs wait for environmentally optimal conditions to start developing. This type of egg is not required to be directly laid on a water body. Picture taken by me.

Larvae hatch from the eggs and remain in the water in order to feed on algae, bacteria and other microorganisms. As for all insects, the larval body has three distinctive sections; head, thorax and abdomen. The larvae live underwater but use their respiratory siphons (except *Anophelines*) for breathing [2]. Larvae moult four times before reaching the pupal stage. Each one of these moulting periods is called an instar. Throughout the last instar and pupal stage, the adult exoskeleton develops internally. Just like egg development, larval development is temperature dependant and can last 6-23 days [2].

After the fourth instar, the larvae grow into pupae, with the pupal stage being the last aquatic stage before emerging into adult mosquitoes. Pupae possess two respiratory trumpets that are used for breathing and, when at rest, the pupae float with the trumpets breaking the water surface for respiration (Figure 1.2-3) [2]. Unlike many other insects, mosquito pupae are very mobile and relatively resistant to desiccation, with mosquito pupae not needing to feed [2]. The pupal stage usually takes 2-3 days, although this can be longer or shorter depending on the environmental temperature [2].

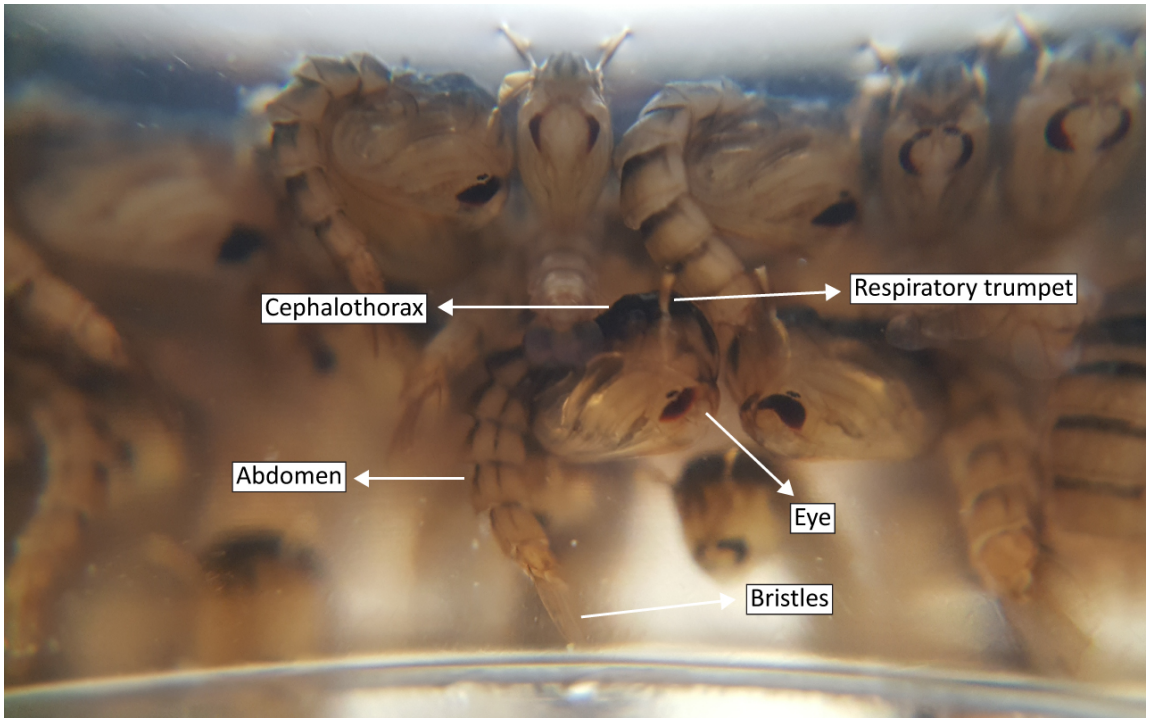


Figure 1.2-3: Close up of an *Ae. aegypti* pupae, collected for transferring into emergence cage. Above the labelled pupa, three pupae can be seen with their trumpets breaking the water surface tension while breathing. The labelled pupa is resurfacing passively after a dive. Picture taken by me.

During the pupal stage, metamorphosis occurs and this is where some of the key processes for adult formation take place. Upon metamorphosis, larval organs that are no longer

required are broken down by a process known as histolysis [2]. The fat body of the larva is transferred to the adult abdomen. The fat body is used as a source of vitellogenines (precursor protein of egg yolks) for autogenous egg formation. At this stage, numerous morphological changes take place [2]. The pupal thorax and head are united into a cephalothorax [2]. The chephalothorax also carries the pupal respiratory trumpets to provide the adult mosquito with oxygen [2]. Both male and female pupae go through these changes and spend the same time in the pupal stage [2]. Although morphologically different under a microscope, the developmental progression from pupa to adult is almost identical for males and females [2].

At the end of the final metamorphosis stage, an adult mosquito emerges. To emerge, the adult mosquito swallows air to expand its thorax. This expansion causes the pupal cuticle to rupture, allowing the adult mosquito to emerge [2]. Once emerged, the cuticle of the mosquito is still soft and starts hardening (sclerotisation) within a few minutes after emerging, allowing adult mosquitoes to fly [2]. Due to the difference in maturity between male and female adult mosquitoes, males emerge 1-2 days before females. During this time (typically 24-48 hours) males reach sexual maturity [2]. Post-eclosing, female mosquitoes are already sexually mature [2]. Although there is a difference between the adult emergence, both spend the same time as pupa, this means the shortening of development time takes place in the larva stage [2]. Between different mosquitoes species, the aquatic stages are quite similar, although very distinctive morphological differences can be observed in adults even by the naked eye.

1.2.1 *Anopheles gambiae* adult

Anopheles adults (Figure 1.2-4) are quite distinctive from the *Aedes* species. *Anopheles* species possess brown pigmentation and can be easily identified by their distinctive linear posture [1]. In their resting position, an Anopheline will form a declining straight line from the end of the abdomen to the tip of the proboscis. *Anopheles* females lay their eggs singly or in small groups and directly on water bodies and *An. gambiae* prefer humans for blood meals [2] prior to oviposition.

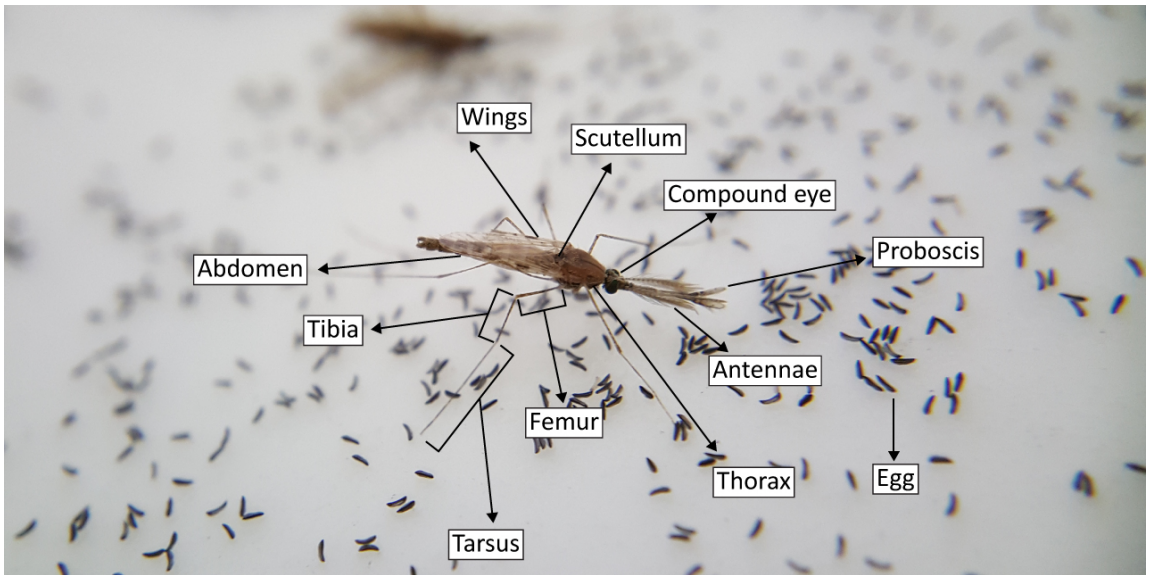


Figure 1.2-4: Close up picture of a mature adult *An. gambiae* male on *An. gambiae* eggs. The distinctive *Anopheles* resting position can be seen from the straight line from the end of the abdomen to the top of the proboscis. Picture taken by me.

1.2.2 *Aedes aegypti* adult

Aedes adults (Figure 1.2-5) are quite distinctive with their black colour and white spotted pattern across their bodies. Compared to the *Anopheles* species, in their resting position they form a roof rather than a straight line. Starting from the end of the abdomen to the scutellum, they form an inclining straight line followed by a declining line from the end of thorax to the tip of the proboscis. *Aedes* females lay their eggs singly or in small groups and can lay eggs on either water bodies or moist soil. Depending on species, female *Aedes* species require either humans and/or livestock for their blood meals prior oviposition.

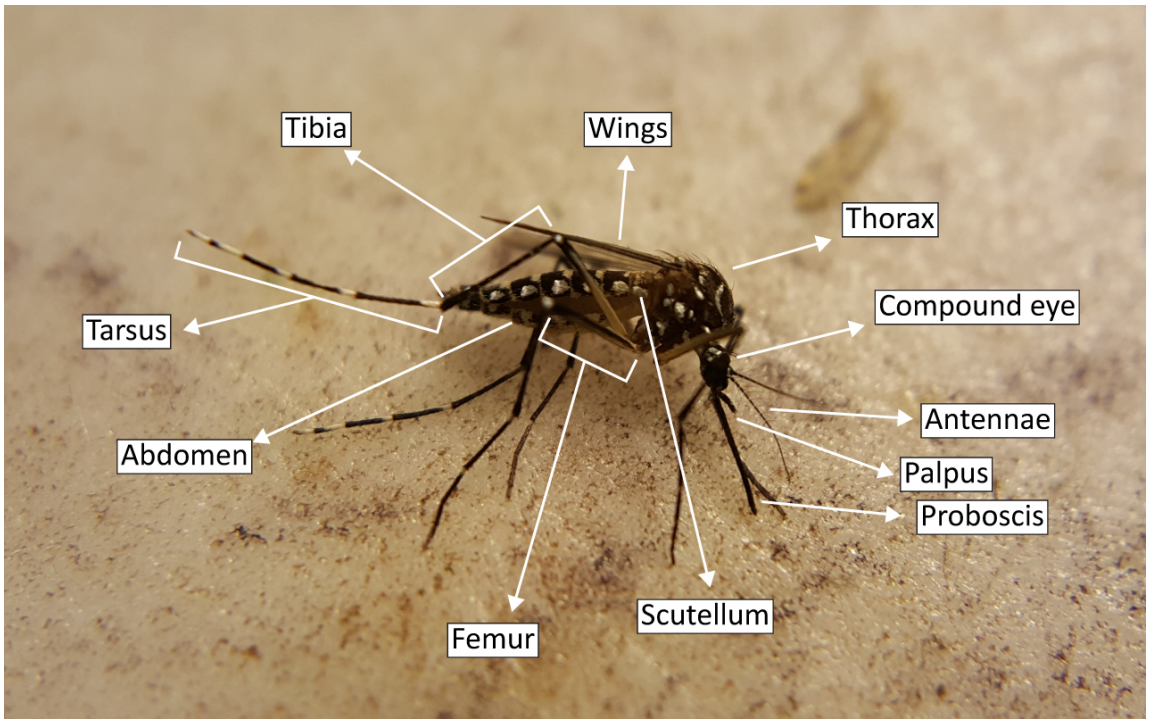


Figure 1.2-5: Close up picture of an adult *Ae. aegypti* male. Distinctive *Aedes* pigmentation of black with white spots can be seen throughout the body and legs. Another distinctive *Aedes* feature is the body forming a roof shape which can be observed at the thorax.

1.3 Insecticides, mechanisms and resistance

1.3.1 Insecticides

Insecticides are compounds used for killing insects. They are often stage-specific such as ovicides and larvicides, which are targeted at eggs and larvae, respectively. The majority of insecticides target either the peripheral nervous system (PNS) or the central nervous system (CNS) [50]–[52]. Table 1.3-1 shows the most common insecticide groups and their targets. Targeting the nervous system is a very effective way of eliminating mosquitoes due to the irreversible effects caused [51], [53].

One of the most well-known insecticides is the organochloride Dichloro-diphenyl-trichloroethane (DDT). It gained popularity as a preventative treatment and has been used to eradicate malaria in western countries such as UK, mainland Europe and USA. However, its use has not been without controversy and whilst DDT was widely used between 1940-1970s, it was eventually banned due to its negative effects on wildlife [51]. Some countries still use DDT, but the majority have converted to more environmentally-friendly insecticides such as the pyrethroids, deltamethrin and permethrin.

Table 1.3-1: Commonly used insecticides, their targets and mode of action.

Insecticide group	Target system	Target	Mode of action	References
Carbamates	CNS	Acetylcholinesterase	Inhibits acetylcholinesterase, causes exhaustion and tetany	[54]
Neonicotinoids	CNS	Nicotinic Acetylcholine receptors	Activates Nicotinic acetylcholine receptors, cause exhaustion	[55]
Organochlorides	PNS	Sodium Channels	Keeps sodium channels open causes seizures	[51]
	CNS	$GABA_A$ receptors	Binds to $GABA_A$ receptors, depresses central nervous system	[51]
Organophosphates	CNS	Acetylcholinesterase	Inhibits acetylcholinesterase, causes exhaustion and tetany	[54]
Pyrethroids	PNS	Sodium Channels	Delay sodium channel closures, causes paralysis	[51]
Ryanoids	Muscular system	Non-voltage gated calcium channels	Binds to ryanoid receptors to lock calcium channels, causes calcium depletion and death.	[56]

CNS, central nervous system; PNS, peripheral nervous system

Currently, the most used insecticides are pyrethroids. These are insecticides derived from the naturally occurring compound pyrethrum. Pyrethroids are favoured due to being less toxic for mammals and the environment. Insecticides of this class target the sodium channels in the PNS.

1.3.2 Common targets, mechanisms and their usage

Most insecticides target the nervous system and their lethal effects are due to the fact that the nervous system is generally irreversibly damaged. Even though an insecticide might target another organ, it is the nervous system that is ultimately affected. Most insecticides target the CNS and PNS although, in recent years, the muscular system has also been targeted.

1.3.2.1 Central nervous system (CNS)

CNS is a widely exploited target in insecticide development. The insect CNS is analogous to the mammalian CNS. Acetylcholine was established to be the neurotransmitter in insect CNS synapses [57], although ɣ-aminobutyric acid (GABA) [58] is also a key neurotransmitter found in the insect CNS. Insecticides targeting the CNS mostly inhibit the acetylcholine esterase, preventing the breakdown of acetylcholine. When this occurs, acetylcholine remains continuously active, causing exhaustion, tetany and eventually death [51].

1.3.2.2 Peripheral nervous system (PNS)

PNS is another widely exploited target for insecticidal activity. PNS comprises all the sensory neurons that are not bundled to the CNS. Within the PNS, the most targeted proteins are of the Na^+ channels [51]. These channels depend on a finely tuned system of Na^+ and K^+ turnover. Insecticides interacting with the Na^+ channels cause delays on the closure of the Na^+ channels [51]. This delay changes the resting potential of the neurons. This perturbation demands more energy, which can often cause paralysis followed by death.

1.3.2.3 Muscular system

Insecticides affecting the muscular system are targeted towards the Ca^{2+} channels in the sarcoplasmic reticulum [59]. When affected by the insecticides, Ca^{2+} ions are rapidly exported to the cytoplasm [59]. This leads to muscle contraction and paralysis [59]. The insecticides in this class have near to zero toxicity to mammals, although a 1995 study revealed an interaction with K^+ channels on mouse models with certain compounds of this class of insecticides [60].

1.3.3 Insecticide resistance and their mechanisms

Insecticide resistance in adults is detected using a response-to-exposure test [61]. An illustration of this test is shown in Figure 1.3-1. This test is conducted using seven chambers (where five are used for testing and two for controls) in an insectary. Each chamber is filled with 20 female mosquitoes and left for an hour. After an hour, dead mosquitoes are replaced. Each chamber is then fitted with an insecticide impregnated paper (oil for control). After exposing the mosquitoes for an hour, impregnated papers are removed. Mosquitoes are then left for 24 hours to recover in the insectary. Post recovery, dead mosquitoes are recorded. The number of dead mosquitoes are summed and expressed as a percentage of all the mosquitoes tested (controls calculated separately). If the mortality of controls is less than or equal to 5%, mortality of test mosquitoes can be interpreted without correction. If control mortality is between 5-20%, the mortality of test mosquitoes needs to be corrected using Abbot's formula [62] to account for the natural mortality rate in the mosquito strain tested. Control mortality higher than 20% indicates the test needs to be discarded and repeated. Interpretation of the test mortality is as follows: higher than or equal to 98% indicates susceptibility, between 97-90% indicates resistance and further tests are required and less than 90% indicates confirmed resistance.

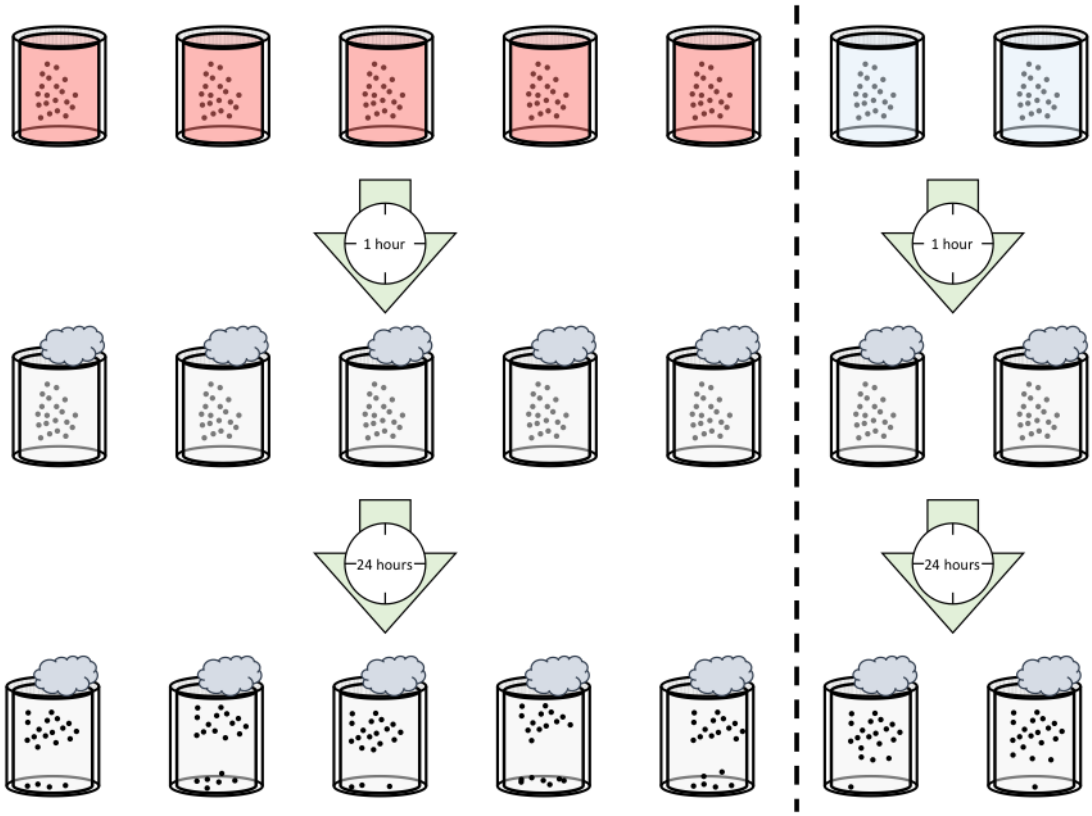


Figure 1.3-1: Determination of insecticide resistance in adults *via* WHO protocol [61]. In an insectary, seven mosquito chambers are set up with 20 female mosquitoes in each. Five chambers (red) are fitted with insecticide impregnated paper and two chambers (blue) are fitted with oil impregnated paper as control. Mosquitoes are exposed to insecticides for one hour and then transferred to clean new chambers with a mesh-screen. A pad of cotton-wool soaked in 10% sugar water solution is then placed on each mesh-screen and the mosquitoes are left to recover for 24 hours. After 24 hours, the number of dead mosquitoes are recorded. The number of dead mosquitoes in the treated chamber are summed up and expressed as a percentage of total number of tested mosquitoes and the same is done for the controls. A mortality rate higher than or equal to 98% indicates susceptibility, between 90-97% indicates possible resistance and requires a repeat of the test and mortality less than 90% indicates confirmed resistance. If mortality in the controls are less than or equal to 5% the test does not require any correction. Between 5-20% requires correction with Abbot's formula. Control mortality higher than 20% requires the discarding and the repeat of the test.

The response-to-exposure test informs on the resistance status. There can be multiple mechanisms causing the resistance, such as behavioural resistance, target site mutation, metabolic resistance and cuticular resistance.

1.3.3.1 Behavioural resistance

Behavioural resistance can be defined as a change in the behaviour of the mosquito affected by pesticides, compared to their regular behaviour [63]. These individuals show a trend to avoid pesticide treated areas and prefer to fly to the pesticide free zones. Thus, avoiding contact with pesticide treated surfaces [63]. Behavioural resistance has been recorded by multiple studies [63]–[66], reporting the recognition of pesticide treatment followed by avoidance of these zone. This has been argued to be one of the reasons for the decline of the

effectiveness of insecticides. However, due to the nature of behavioural studies, designing and executing experiments to test this hypothesis has proven to be a major challenge.

1.3.3.2 Target site mutation

Another resistance mechanism that arises through overuse of insecticides is target site mutation. In target site mutation resistance, a point mutation in the targeted enzyme causes the insecticide less favourable or impossible to bind to the targeted enzyme. This causes the insecticides to be excreted as waste material [67]–[71].

1.3.3.3 Metabolic resistance

Metabolic resistance is caused by the fast metabolic breakdown of the insecticides. In this resistance mechanism, the insecticides taken up by the mosquito are rapidly broken down to non-toxic compounds and excreted. Although this may not be as efficient a mechanism as target site mutation (where the insecticide cannot activate the lethal mechanism it is designed to), with metabolic resistance the lethal concentration is not reached since the insecticides are rapidly broken down. Extreme dosage use can still be lethal, although, at these extreme doses the insecticides become toxic to other organisms where the insecticide applied as well [70]–[72].

1.3.3.4 Cuticular resistance

The cuticular layer, also known as the exoskeleton, is the outermost layer of an insect. The majority of insecticides targeting adult mosquitoes reach their target by penetrating this layer. The cuticular layer is comprised of epicuticle, exocuticle and endocuticle.

Epicuticle forms the outer most layer of the cuticle and is typically 1-4 µm in thickness [73]. A detailed composition of this layer is difficult to elucidate but it is known to contain sclerotized proteins, lipoproteins, lipids, waxes and a shellac-like substance (similar to resin) [73]. This layer forms the outer most barrier to the cuticle. Underneath epicuticle is the procuticle which is formed by exocuticle (top) and endo cuticle (bottom) [73]. Exocuticle is a hard and sclerotized layer containing chitin and proteins. Endocuticle is mainly composed of hydrocarbons, proteins, lipids, wax esters and free fatty acids [73]. Hydrocarbon types that can be observed in this layer are n-alkanes, methyl-branched alkanes and unsaturated hydrocarbons [72]. Endocuticle also contains chitin and proteins although due to lesser sclerotization it is softer and more flexible in comparison [73].

Epidermis is a layer comprised of epidermal cells underlying the cuticle. Insecticides penetrating through this layer reaches the basal lamina which have pores facilitation access to haemolymph [73].

Cuticular resistance refers to the changes in thickness and/or composition of the [74], [75]. Figure 1.3-2 shows an illustration of how different cuticular layer structures can affect the uptake of insecticides.

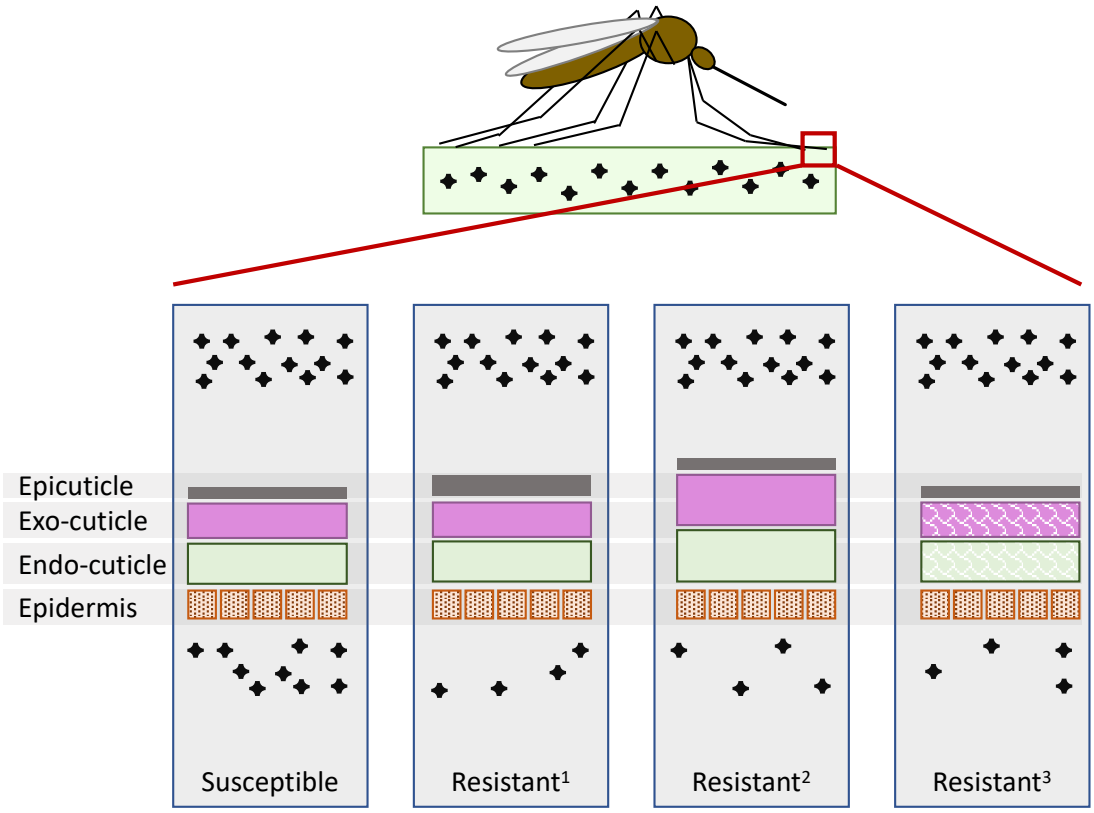


Figure 1.3-2: Insecticide penetration efficacy on different cuticular layers. In the susceptible cuticle, insecticide penetrates as intended and is absorbed through the epidermis. In the altered cuticle layers Resistant^1 (thicker epicuticle), Resistant^2 (thicker endo-cuticle and exo-cuticle) and Resistant^3 (change in cuticular layer composition), insecticide resistance is considerably greater than the susceptible cuticle (adapted from [74]).

The cuticular layer of a susceptible species is more permeable to insecticides. When the cuticular layer is altered by means of thickening of layers and/or alteration to the composition of the layers, the permeability to insecticides is reduced [76]. In this case, if a lethal dose insecticide was applied, only a small fraction of the insecticide penetrates the cuticular layer and mosquito survives the application since lethal insecticide levels are not reached.

In current insecticide resistance research, behavioural resistance is still treated with caution. Showing evidence of insecticide avoidance has proved to be a major challenge. Both metabolic resistance and target site mutation are very well studied and characterised with a wide repertoire of studies covered in numerous reviews [68], [70]. Cuticular resistance is a more recent discovery and as such requires further investigation [75], [77].

1.4 Cuticular Hydrocarbons (CHC)

1.4.1 CHC in mosquitoes and their functions

Mosquito integument is formed from three elements: cuticle as the outer layer, epidermis, and basement membrane [1]. The integument functions both as a protective layer for the organs and as a rigid exoskeleton for the muscles to bind to [1]. This integument also serves as a barrier for water loss [78]. The outermost layer of the integument is the cuticle and it is first line of the mosquito's defence against both living threats (such as fungi, bacteria and parasites) and non-living threats (such as insecticides). In some insects, the cuticle also serves an instrument of communication which facilitates the selection of mating partners [79], communication of caste [80], swarm [81], and autotoxicity [82]. Evidence of similar functions been shown in mosquitoes [83]–[85] although overall their function is known to a lesser extent.

The outermost section of the cuticular layer comprises a variety of long chain alkanes, branched alkanes, alkenes, waxes and esters. Chung and Carroll [86] have shown CHCs with higher melting temperatures (T^m) directly correlate with waterproofing properties, whereas lower T_m CHCs are correlated with information content (as contact pheromones) (Figure 1.4-1).

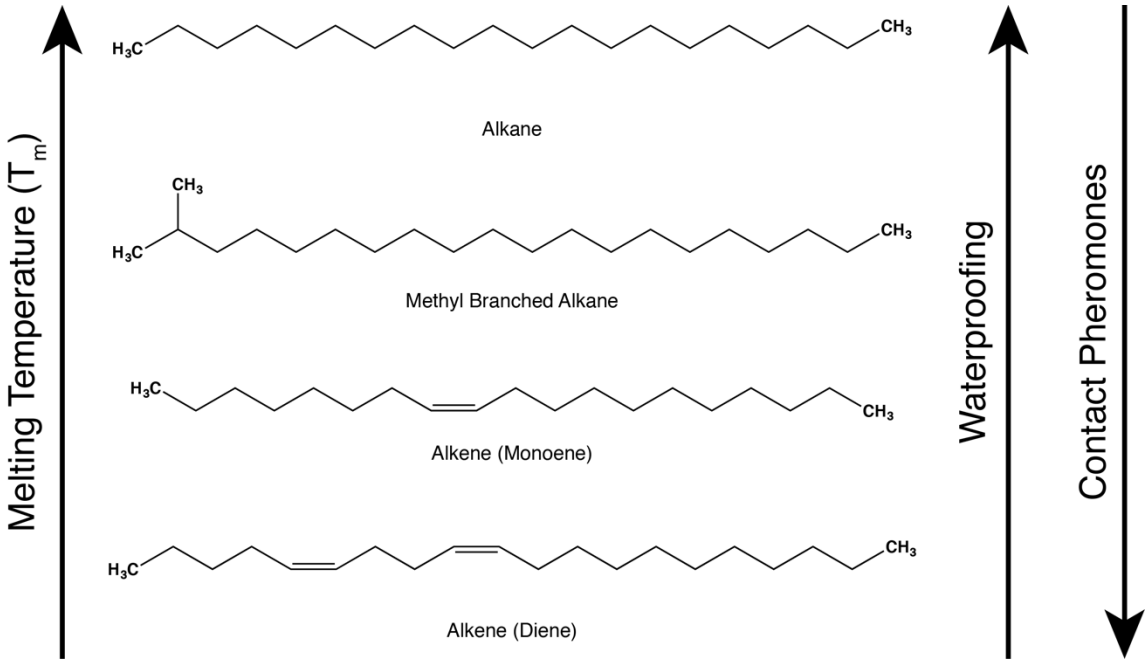


Figure 1.4-1: Examples of typical mosquito CHCs and their function in relation to T_m. CHCs are typically made of 16-37 carbons (adapted from Chung & Carroll, 2015).

1.4.2 CHC in insecticide resistance

Adult mosquitoes can come into contact with insecticides in one of two ways. Either the insecticide molecules in the air land on the mosquito's body or the mosquito lands on an insecticide-impregnated surface and insecticide is contacted through the tarsus (lower section of the leg, Figure 1.2-5). Between the two, the latter is the most common form of exposure to insecticides due to their application regimen. Upon interaction, insecticides penetrate the cuticular layer down to the epidermis and eventually make contact with the target proteins. In mosquitoes with cuticular resistance, this process is either slowed down and/or altered.

A 2012 study by Qiu *et al* [87] investigated CHC biosynthesis in insects. In their search for a candidate aldehyde oxidative decarbonylase cytochrome P450 (CYP), they identified CYP4G subfamily as being present in every insect genome. From microarray data Cyp4g1 was identified as the most highly expressed of all CYP genes in *D. melanogaster*. Furthermore, they identified Cyp4g2 as the closest homologue in *Musca domestica* which was used to show the decarbonylase activity *in vitro* [87].

A study conducted in 2016, assessed insecticide penetration in *An. gambiae* using ^{14}C deltamethrin [75]. The study showed that resistant species had slower uptake of deltamethrin. In the same study, this result was followed up by electron microscopy which

measured the thickness of the cuticular layer from a femur cross-section. It was found that the resistant species had a significantly thicker epicuticle (2.13 ± 0.32 μm) than the susceptible species (1.873 ± 0.30 μm). In the same study they identified Cyp4g16 and Cyp4g17 as potential paralogues of Cyp4g1 and Cyp4g2 [75].

1.4.3 CHC Biosynthesis

CHC composition varies between different insects. For example, in *Drosophila* CHCs differ between the species and within *Drosophila melanogaster* different CHC profiles can be seen between males and females [86]. Despite numerous different configurations, insect CHC biosynthesis is very well conserved [88]. CHCs are synthesised in the oenocytes which are secretory cells that are characteristic of insects. Between different species, the oenocytes can be found in different locations; in mosquitoes they are mostly found in the abdomen. Oenocytes work together with the fat body to secrete components of the cuticular layer, which is followed by transportation to the cuticle *via* lipophorins [88].

CHC synthesis is a diverse set of reactions (Figure 1.4-2) with a wide range of end products. Within the oenocytes, CHCs are synthesised in the endoplasmic reticulum. It has been shown that acetate is a common precursor of CHC biosynthesis *via* labelled acetate studies in *D. melanogaster* [89]. The proportion of labelled acetate derived carbons was found to be equal across all HCs, regardless of the HC type.

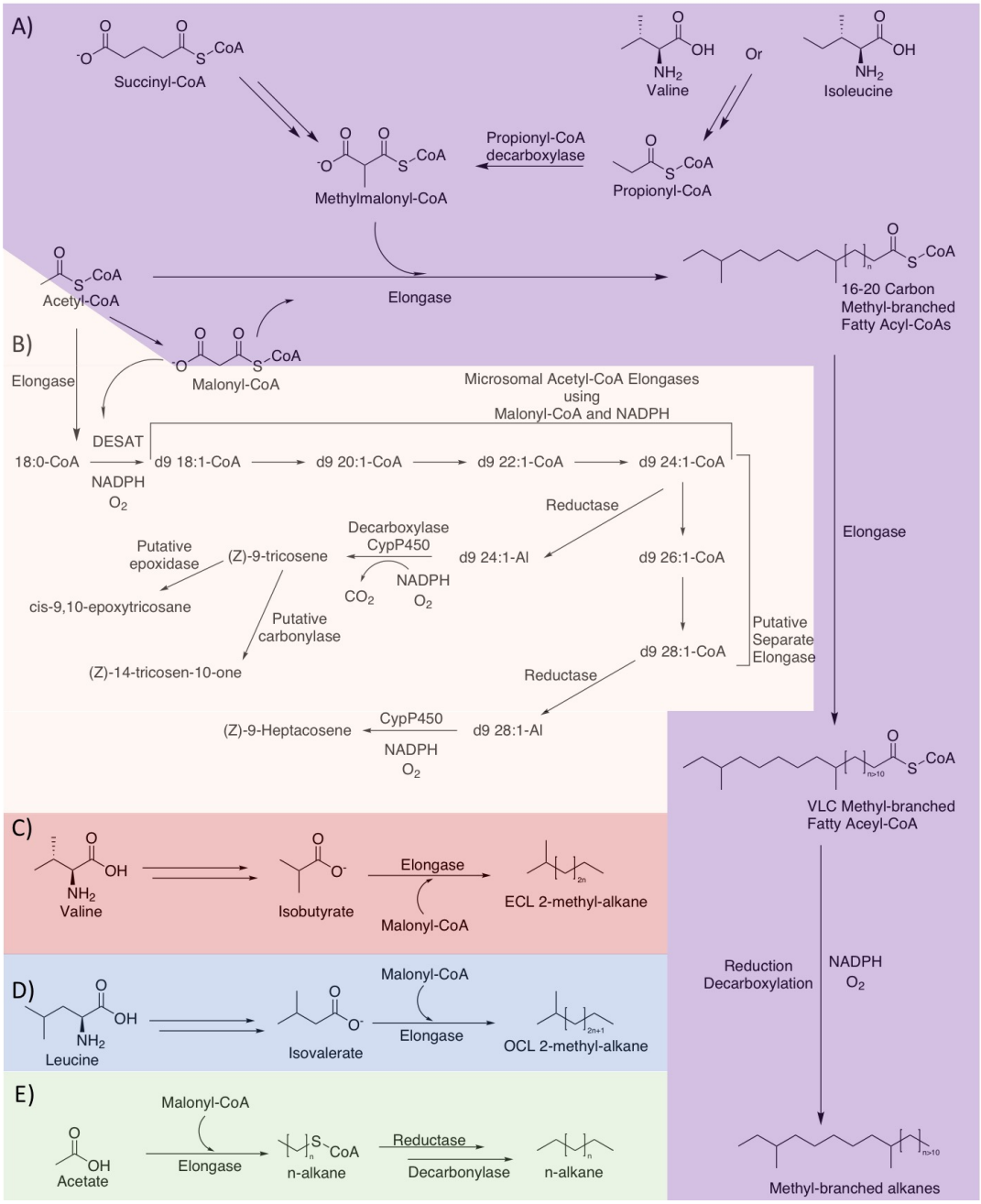


Figure 1.4-2: General CHC biosynthesis pathway in insects. In each pathway, malonyl-CoA (non-branched) and/or methyl malonyl-CoA (methyl branched) are used for hydrocarbon elongation. Double arrows represents omitted multiple steps. A) Biosynthesis of methyl-branched alkanes (purple) from valine, isoleucine, succinyl-CoA, and acetyl-CoA. 16-20 carbon methyl-branched fatty acyl-CoA are elongated to very long chain (VLC) methyl-branched fatty acetyl-CoA. Prior to producing the final product, CoA is cleaved off by a reductase and finally the carboxyl group is decarboxylated *via* CypP450. B) Shows the production of alkenes (yellow) starting with an Acetyl-CoA as a head group. The HC chain is elongated using malonyl-CoA and elongase with each elongation step requiring NADPH for the elongase to catalyse the reaction. The final product, (Z)-9-tricosane, is decarboxylased *via* CypP450. (Z)-9-Tricosane can also be used as a starting material for further alkene biosynthesis. DESAT: desaturase, Hyd: hydrocarbon, Al: aldehyde. C) Biosynthesis of even chain length (ECL) 2-methyl branched alkanes (red) $_{\gamma}CH_3$ of valine becomes the 2-methyl branch in the final product where elongation is done by addition of malonyl-CoA. D) Biosynthesis of odd chain length (OCL) 2-methyl branched alkanes (blue) $_{\Delta}CH_3$ of leucine becomes the 2-methyl branch in the final product where elongation is done by addition of malonyl-CoA. E) Biosynthesis of n-alkanes (green) acetate is used to produce appropriate length fatty acyl-CoA which are then elongated using malonyl-CoA. Figure adapted from (Howard & Blomquist, 2005).

In order to synthesise n-alkanes, appropriate length fatty acyl-CoAs are elongated using acetate in the form of malonyl-CoA [90]. Acetate is converted acetyl-CoA *via* acetyl-CoA synthase [91] which is then carboxylated to malonyl-CoA by acetyl-CoA carboxylase [92]. Post elongation, a decarbonylation step catalysed by a CYP yields the final product which has one less carbon than the elongated precursor [90]. Branched alkanes are synthesised in a manner similar to n-alkanes. In order to create the branching, malonyl-CoA is substituted with branched amino acid (valine, leucine and isoleucine) derivative methylmalonyl-CoA [93]. Biosynthesis of even chain length 2-methylalkanes starts by converting valine to isobutyrate and for odd chain length leucine to isovalerate [94].

The biosynthesis of alkenes follows the same process as that of the n-alkanes and methyl-branched alkanes, with the addition of fatty acyl-CoA desaturases. The majority of insects have the enzyme Δ^9-desaturase hence insect alkenes predominantly have the double bond at the 9-position [88]. Studies on the desaturases in *D. melanogaster* characterised two proteins, desat1 and desat2, which take part in the biosynthesis of 7,11-double bonds and 5,9-dienes, respectively [88].

Prior to the final step, HC-CoAs are reduced to HC-aldehydes *via* reductases. Studies on *D. melanogaster* and *An. gambiae* show the decarbonylation of the HC-aldehyde *via* CYP decarbonlyases [75], [86]. Decarbonylation enzymes are highly conserved between species and are found as paralogues [95]. Enzymes catalysing the previous steps of HC biosynthesis are found as multicopy and specific isoforms are expressed in oenocytes [96]. The common decarbonylation step in the CHC synthesis pathway is an ideal target to probe CHC production using metabolomics.

1.5 Cytochrome P450 (CYP)

CYPs are a heme-binding, electron transferring, superfamily of proteins. CYP contain a wide range of families of enzymes which take part in numerous reactions, ranging from drug metabolism & detoxification to fatty acid biosynthesis. CYPs can be found throughout the domains of archaea, eukarya, bacteria and viruses [95], [97]. Although they have a wide range of substrates, CYPs have an overall conserved fold which gives this superfamily its conserved mechanism for a wide range of reactions [95]. The majority of CYPs catalyse monooxygenation [97] reactions and can be generalised as:

$$RH + O_2 + NADPH + H^+ \rightarrow ROH + H_2O + NADPH^+$$

In this reaction, a hydroxyl group is incorporated to R following donation of an H^+ from NADPH to the monooxygenase enzyme. After the monooxygenation, remaining oxygen is reduced to H_2O. CYP enzymes require a redox partner and are generally divided into two major classes. Class I CYPs comprise mitochondrial and bacterial P450s that use two unique redox partners involving an iron-sulphur protein and a flavin-containing reductase making a three component system [95], [98]. Class II CYPs are microsomal monooxygenases and partner with NADPH-cytochrome P450s oxidoreductase (POR) for supply of electrons required for monooxygenation, thereby creating a two-component system [97], [98]. The overall structure of CYPs is quite conserved. When different CYP structures compared, majority of the differences come from ligand binding sites. In contrast the heme-binding sites are highly conserved [95]. In eukaryotes, CYPs are generally bound to the endoplasmic reticulum membrane [95].

1.5.1 CYPs of mosquitoes

Microarray and transcriptome studies conducted on vector insects have shown overexpression of CYP4, CYP6, CYP9, CYP12, CYP305, CYP307, CYP314 and CYP325 families in insecticide-resistant populations [99]–[105]. Genomic and transcriptomics studies on CYPs have shown that CYP4G genes are insect-specific [106]. In 2011, a study on the mosquito species *Culex quinquefasciatus* identified a total of four families in the completed genome: CYP2, CYP3, CYP4 and mitochondrial [107].

The majority of the studies on resistant mosquitoes mainly focus on target site mutations and metabolic resistance [52]. In these studies, members of CYP6 and CYP9 families are usually flagged as key components of these resistance types due to their roles in detoxification mechanisms [108]–[111].

In recent years however, there has been more interest in the cuticular resistance. Studies focused on genomics and transcriptomics have highlighted the importance of the CYPs associated with cuticular layer biosynthesis [74], [75], [87] (Section 1.4.1).

The CYP4 family catalyses hydrocarbon biosynthesis and metabolism. Enzymes of this family mainly act as monooxygenases, but there are members that act as decarbonylases [95] (Figure 1.5-1). Several studies have shown overexpression of CYP4 enzymes in mosquito

species with cuticular resistance along with some target site mutation resistance [75], [87], [112]. In 2012, a study conducted on *D. melanogaster* and *Musca domestica* [87] targeted two CYPs active in the final step of CHC biosynthesis (~C21-C37+ length). In this study, Qiu *et al* used a Gal4-UAS system to create a Cyp4g1 knock-down line of *D. melanogaster* [87]. These knock-downs resulted in a reduction of CHCs and high mortality rates in post larval stages [87]. In the same study, the orthologue Cyp4g2 (*M. domestica*) was expressed as a recombinant protein in *Saccharomyces cerevisiae* and its decarbonylase activity was determined *in vitro* [87]. Similar results regarding the number of deaths in knock-down strains were recorded by other studies [113]–[115]. Qiu *et al* proposed the use of the CYP4 family of enzymes as a target for insecticide resistance [87].

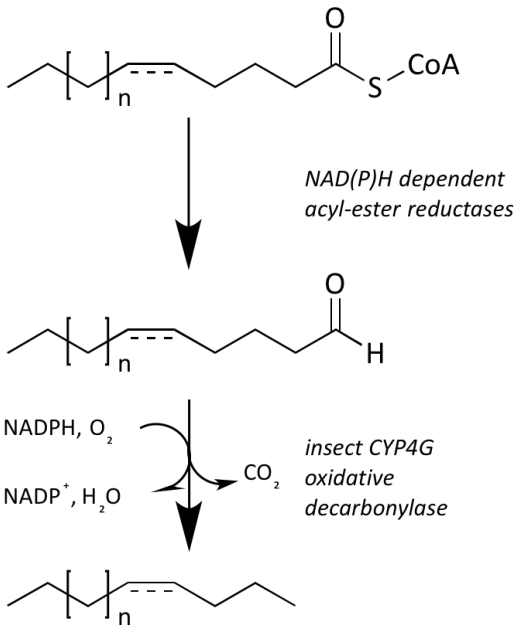


Figure 1.5-1: Generalised form of CHC production through CYP-P450, where a decarbonylation step taking place. The majority of CYP-P450s perform a decarboxylation (adapted from Qiu et al, 2012).

1.5.2 Cyp4g16 and Cyp4g17

Cyp4g16 and Cyp4g17 are the closest *An. gambiae* paralogues of the previously mentioned Cyp4g1 and Cyp4g2 enzymes [75]. Currently, there are no deposited structures of Cyp4g16 and Cyp4g17. A 2016 study by Balabanidou *et al* has demonstrated that Cyp4g16 and Cyp4g17 are two critical CYPs in the CHC pathway and are found to be up-regulated in insecticide resistant species [75]. Both Cyp4g16 and Cyp4g17 are localised to the oenocytes, with immunohistochemical experiments showing Cyp4g16 to be localised to the plasma membrane and Cyp4g17 to around the cell nucleus, the latter suggesting an association with the endoplasmic reticulum [75].

Studies on CHCs and their related enzymes typically take a more direct approach, such as a labelled trace study or expression of the enzyme to demonstrate its properties *in vitro*. However, these approaches may not be suitable for every case. For example, the expression of Cyp4g16 and Cyp4g17 in the work by Balabanidou *et al* was only successful for Cyp4g16 [75]. The function of Cyp4g17 was predicted due to their high similarity [75]. Although expression of Cyp4g17 was unsuccessful, it is possible to create knock-down lines of this enzyme. A different approach to such a case is the metabolic profiling of these knock-down lines in order to investigate CHC biosynthesis precursors. Such an approach would be ideal with metabolomics methods.

1.6 Metabolomics

1.6.1 Metabolomics in science

Metabolites are small biological molecules, typically less than 1500 Da which are often intermediates or products of biological processes [116]. Metabolites also have roles in the regulation of certain processes as signals and/or checkpoints [117]. By measuring metabolic perturbations, information on biological processes can be recorded or inferred.

Metabolomics is an application of analytical chemistry. Employing analytical methods to determine the abundance, relative or absolute, of a molecule predates the term metabolomics. Emergence of the field of metabolomics coincided with advances made in spectroscopic, mass spectrometry and chromatographic methods, allowing samples to be analysed in a high sensitivity, high resolution and high throughput manner [117]. Since its emergence in the early 1970s, metabolomics has been applied to many areas of biological research including microbiology, oncology, agriculture, animal studies and insect biology. Taking its strength from identification and quantification of a range of metabolites simultaneously, metabolomics applications are varied and include diagnostics, disease progression, disease mechanism elucidation, biomarker discovery, personalised medicine, cell biology and experimental condition comparisons.

1.6.2 Core concept

Metabolomics is the study of small molecules in a system *via* quantification, be it relative or absolute. By quantifying metabolite levels, a metabolic profile (chemical phenotype) is built. This profile is then compared to another profile built under a different experimental condition [118] for example wild type v knock-down or treated v untreated. This comparison

allows investigation of processes governing the biological system in question. Complementary to other "-omics" techniques (genomic, transcriptomics and proteomics) (Figure 1.6-1), metabolomics focuses on the precursor and product material in biochemical reactions [118]. Ultimately, the metabolites reflect the bio-processes that the system has achieved at any time point, building upon the information from potential processes reported in proteins and transcripts. Furthermore, metabolomics can infer information on the most predominant processes, which is not possible from protein levels alone.

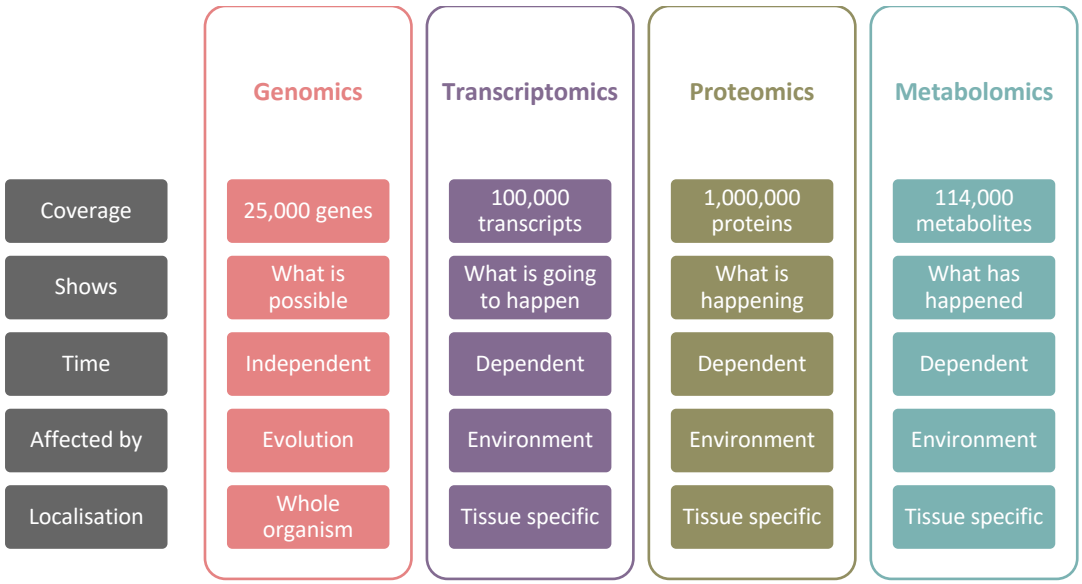


Figure 1.6-1: General properties of all -omics techniques. Comparing the human coverage of each technique, metabolomics has a rather large set with 114,000 metabolites compared to the estimated 25,000 genes, 100,000 transcripts and 1,000,000 proteins [117], [119].

In simple terms, genomics creates the potential map of an organism, showing all the pathways that could be taken in different conditions [118]. Transcriptomics uses the genomics data in order to identify the genomic instructions being executed [118]. The degree of execution is reflected in terms of the number of each transcript expressed. Transcripts provide the genetic code that directs production of proteins, which in turn catalyse the chemical reactions producing or regulating metabolites. Proteomics identifies and quantifies proteins in a system which yields information on active processes in the system [118]. Every chemical reaction starts with substrate/precursor molecules and ends with product molecules. Metabolomics monitors the abundance of these molecules, thereby reporting on the processes that have just occurred [118].

1.6.3 Approaches in metabolomics studies

Metabolomics research typically follows one of two approaches; targeted or non-targeted [120]. A targeted approach is preferable for the measurement of a small and specific set of metabolites which are expected to change between experimental conditions. Studies of a targeted nature are inherently easier to quantify, although, making assumptions on the whole system beyond the measured metabolites is considerably harder or simply not feasible in some cases. Another downside of targeted metabolomics is that these studies may 'miss' the knock-on effects of seemingly disparate processes [120]. Non-targeted metabolomics, sometimes called "discovery" metabolomics, aims to identify and measure as many metabolites as possible. This approach is ideal when comparing experimental conditions and/or identifying metabolic changes. However, it should be noted that mapping metabolomics information back to biological processes is only as good as the information and databases available on the organism of interest [121].

The choice of approach is dependent on the research question and compromises between metabolite coverage and precision. A system-wide approach such as investigating all the metabolites in a cell culture or a urine sample will give rise to a rich and complex metabolite profile, but quantification will be challenging. However, looking for several specific metabolites would be easier to quantify but as the coverage is very narrow, certain processes would not be reported.

1.6.4 Metabolomics workflow

For an effective metabolomics workflow (Figure 1.6-2), careful consideration of the biological rationale is essential, with a clear understanding of what information metabolomics will generate [118]. This is followed by careful experimental design which must consider and mitigate specific limitations of the study, as metabolites susceptible to any change in experimental design which will severely effect study outcomes. Due to the statistical nature of the analysis, consistency is a very critical part of metabolomics studies. Following the biological experiments, samples are carefully collected and prepared for the required analytical platform. Depending on the experimental design, sample preparation might include a chemical extraction [118]. Following sample preparation, data is acquired using the appropriate platform. Data is then checked for quality; this step ensures the data was acquired to the desired standard and helps with identification of required repeats. Prior to statistical analysis, the data is checked for outliers and artefact data points. Cleaned data is

then subjected to robust statistical analysis. Finally, the important findings identified through statistical methods are contextualised biologically. Attributing metabolite changes to biological processes is often the most challenging part of the metabolomics workflow [116].

Figure 1.6-2: Typical Metabolomics workflow. Alternating dashed line represents the iterative step where samples failing the quality control are repeated. Uniformly dashed line represents optional routes, which might be required depending on the experimental design and research question.

1.6.5 Techniques in metabolomics - advantages and disadvantages

Metabolomics methods can be categorised under two general groups: analytical and statistical. The analytical techniques utilised in metabolomics are some of the most common tools used in analytical chemistry: mass spectrometry (MS) and solution nuclear magnetic resonance spectroscopy (NMR) [122]. Technological advancements in these analytical tools have allowed their usage in a high throughput manner, which is a key feature in metabolomics. After acquiring data with the analytical platform of choice, a final dataset table is generated which is then analysed *via* a series of statistical methods. In this section, the advantages and disadvantages of both analytical and statistical techniques will be discussed.

1.6.5.1 Nuclear Magnetic Resonance

NMR is a cornerstone analytical tool with a wide range of applications that includes compound quantification, purity assessment, structural elucidation, and binding constant calculations to name but a few. NMR spectroscopy, like any other spectroscopy, measures

the interactions between matter and electromagnetic radiation. NMR can be separated in to two main types, namely solution NMR and solid-state NMR. Both types of NMR operate under the same physical principles in terms of interaction of matter and electromagnetic radiation. Since only solution NMR was used in this project, only solution NMR will be explained and discussed, and any instance of NMR hereafter refers to solution NMR.

NMR exploits an intrinsic property of atomic nuclei called spin [123]. Spins are an intrinsic nuclear angular momentum that cause the nuclei to behave like magnets. As a consequence, nuclei with a net spin adopt a preferred orientation when exposed to an external magnetic field. With any nuclei there are 3 possible ways the spin can be quantified:

1) The number of neutrons and protons are both even resulting in no spin.
2) The number of neutrons plus protons is odd resulting in half integer spin (i.e. 1/2 , 3/2, 5/2).
3) The number of neutrons and protons are both odd resulting in integer spin (i.e. 1, 2, 3).

The quantification of spin determines energy levels that a nucleus can take in an applied magnetic field [123]. This is determined by the equation $2I + 1$, where I is the net spin of the nuclei [123]. For example, the most typical NMR active nuclei Hydrogen (or as typically referred to, a proton) has a net spin of ½, which means in an applied magnetic field it has two energy states, high energy state (β) and low energy state (α) [123]. Nuclei with half integer spin other than ½ will have more than 2 energy levels and the difference between the energy levels is directly proportional to the strength of the magnetic field. For simplicity, half integer spin of ½ will be explained in this section, although the same concept is applicable to different numbers of energy levels [123].

As previously mentioned, a proton (net spin of ½) will adopt a preferred orientation in an external magnetic field. This is determined by either the high or low energy state. A proton's preferred orientation in a magnetic field is either in the direction of the magnetic field (aligned, β state) or in the opposite direction (against, α state), but not parallel to the magnetic field [123]. However, it is important to remember that not all nuclei in a sample will show this behaviour in a magnetic field. In a magnetic field of 2.34 Tesla, which is the same as a 100 MHz NMR magnet, only 6 in every 1,000,000 nuclei are in the α state, with

even less nuclei populating the higher energy β state. The remaining nuclei that are not in either the α or β states are randomly oriented in the magnetic field [123]. When all of the orientations of the nuclei in a sample are summed, the randomly oriented nuclei sum up to zero and the net bulk magnetization is therefore given by the sum of the nuclei in the α and β states [123]. This low proportion of energy level difference is enough to make measurements, but also gives NMR its inherent low sensitivity [123]. The number of α and β states are directly proportional to the strength of the magnetic field, meaning a higher magnetic field will yield a higher energy level difference thus increasing the sensitivity of the measurement.

The relationship between the preferred orientation of a nuclei and the energy states lays the ground for NMR measurements. As discussed, when a proton is in a high magnetic field, typically annotated as B_0, it will either be in a high energy state (β) and align with the external magnetic field or be in a low energy state (α) where it is against the external magnetic field. However, this ground state of nuclei in the α and β states is not static, but rather it precesses at a certain frequency, also known as Larmor frequency, around the magnetic field axis. Furthermore, these α and β states can be manipulated by applying radio frequencies which then give rise to resonance energy at a frequency intrinsically unique to that atom [123].

The radio frequencies used for pulsing a sample are in the MHz range, the same frequency range as used for car radios. This is where the resonance in NMR is critical. Resonance can be defined as the amplification of a measured event due to the matching of frequencies. As such, when a sample is pulsed with radio frequencies for measurement, it is matched, or in resonance, with the Larmor frequency [123]. This burst of in resonance electromagnetic pulse through a coil around the sample creates a temporary magnetic field (B_1), which is perpendicular to the static magnetic field B_0. This causes the precessing nuclei to become aligned along the new magnetic field axis B_1 [123]. Measurements in NMR are taken by the same coil around the sample. The signal recorded will be amplified due to the in-resonance pulse applied prior to the acquisition of the resonance signal.

The recorded signal in NMR is a free induction decay (FID) and it is the superimposition of all acquired resonance signals [123]. The FID has a duration, typically up to 5 seconds depending on the sample, and an intensity. For the analysis of such complex signals, Fourier transformation is ideal. Fourier transformation is a mathematical operation where the time

domain is transformed to the frequency domain [123]. A Fourier transformed FID results in a spectrum where the x-axis is frequency and the y-axis is intensity and the signatures of all the detected nuclei can be observed [123].

The intrinsic resonance frequency of each atom is known as its chemical shift and is measured in Hz. Since the precession is directly proportional to the magnetic field, two molecules measured on two magnets with different magnetic fields, e.g. 600 MHz and 800 MHz, would not have the same chemical shift [123]. In order to make NMR spectra comparable between different magnetic fields, the chemical shift needs to be normalised to the magnetic field. For example, a signal measured at 600 Hz on a 600 MHz magnet would be 600 Hz/ 600 MHz, yielding 1 x 10^{-6} or 1 part per million (ppm). Using higher magnetic fields comes with two advantages, increase of sensitivity and resolution. Earlier it was mentioned how higher magnetic field increases the population difference observed in bulk magnetisation hence, increasing sensitivity [123]. Resolution enhancement can easily be explained with the ppm scale. Two spectra acquired at 600 MHz and 800 MHz would have the exact same ppm scale. Taking a portion of this scale as an example (0-1 ppm), spectra from the 600 MHz magnet would have points from 0-600 Hz to define 1 ppm while the 800 MHz magnet would have points from 0-800 [123]. An illustration of resolution enhancement with higher magnetic fields can be seen in Figure 1.6-3.

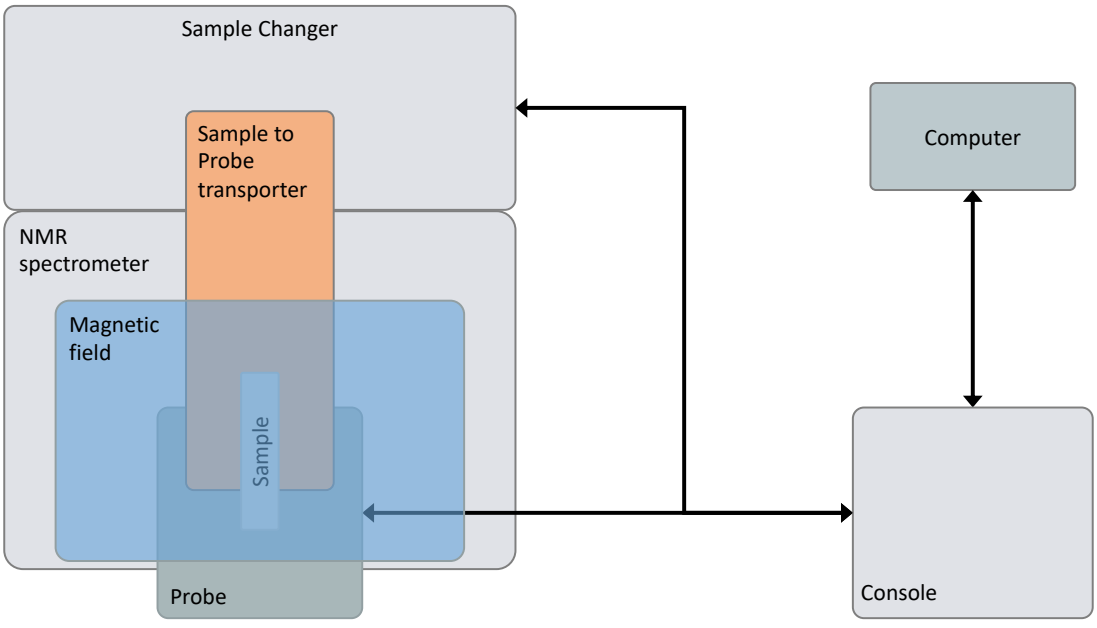


Figure 1.6-3: Basic schematic of a NMR spectrometer. The NMR sample vial is placed onto the probe that resides in a magnetic field. When the sample is placed in a strong magnetic field, net magnetisation of the spins either

align with or against the direction of the magnetic field. Using RF pulses generated from the probe, these net magnetisations are manipulated. These changes are transferred by the probe to the console for digitisation of the data followed by transferring to a computer for analysis.

The chemical shifts obtained from NMR are related to the chemical structure of the molecule the atom is part of [123]. A methyl group resonance which is typically 0.5-1 ppm is distinctly different than the resonance of an aromatic ring, typically 6.5-8 ppm. Certain molecular structures will cause signals to split in a systematic fashion. A split signal is called a multiplet and the multiplicity of the signal is given by the rule $n + 1$, where n is the number of neighbouring ^{1}H environments (see the origins of the signals in Figure 1.6-4). Additionally, the intensity pattern also follows a systemic rule based on Pascal's triangle [124]. Furthermore, the intensity of the resonance peak is proportional to the abundance of the molecule and it is this particular property makes NMR more amenable for quantification.

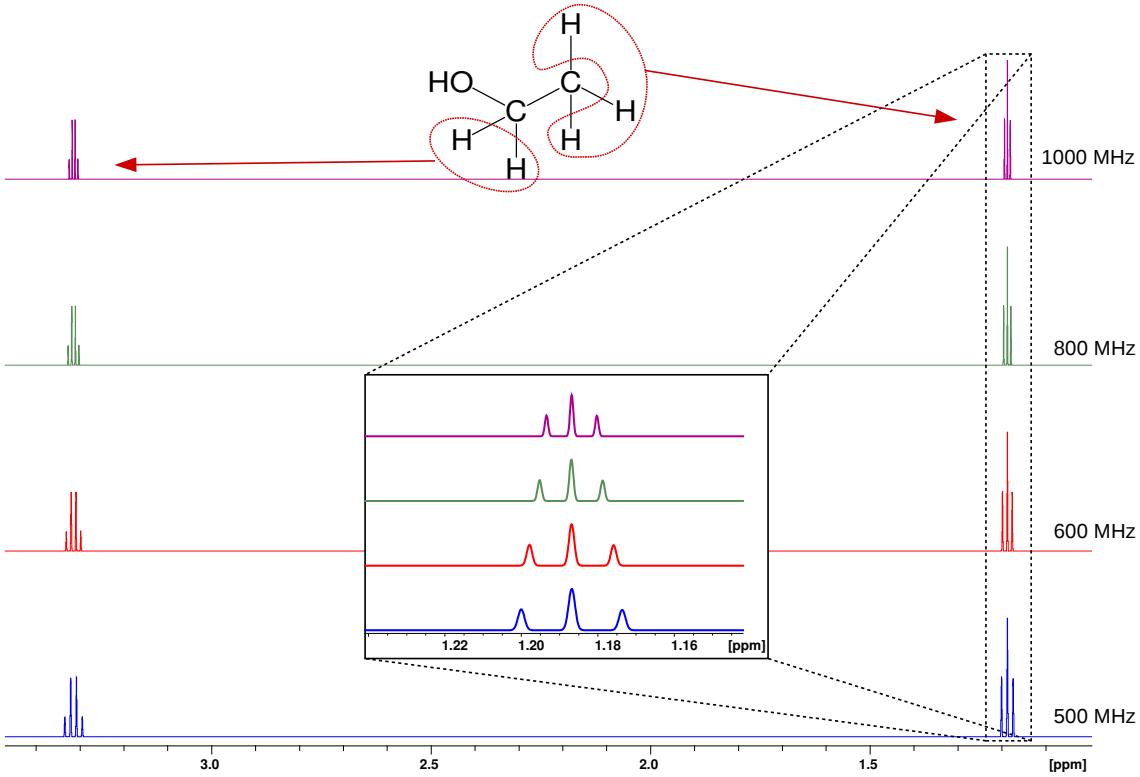


Figure 1.6-4: Stacked spectra of simulated ^{1}H of ethanol under magnetic fields of 500 MHz, 600 MHz, 800 MHz and 1000 MHz. All spectra show the range of 1 ppm to 4 ppm where the signals from CH_2 and CH_3 are observed. Higher resolution provided by higher magnetic fields can be seen from the narrowing of the multiplet peaks as the magnetic field increases. Spectra were simulated using nmrdb [125]–[127].

In biological sciences, molecules of interest are typically composed of a combination of H, C and N atoms. All three elements have a NMR active (spin ½) isotope: ^{1}H, ^{13}C, and ^{15}N. Due to its high natural abundance (99.99%), in addition to being a key element in most organic

molecules, proton NMR, or ^{1}H NMR, is the most commonly used NMR method [123]. However, there is one major disadvantage with ^{1}H NMR, its narrow ppm scale. All ^{1}H signals are typically observed between 0-12 ppm. For comparison, the ^{13}C NMR scale is typically between 0-200 ppm. The high numbers of ^{1}H atoms in organic molecules, compounded with the narrow scale at which they resonate, means ^{1}H NMR spectra of complex mixtures are often overcrowded and highly overlapped [128]. This makes the identification of metabolic signatures a challenge. Even though splitting and intensity patterns can be used to deconvolute some signals, for complicated mixtures this can be very demanding. Improving the resolution of signals by using high magnetic fields can be of great aid in the identification of molecules with very close chemical shifts. A typical NMR metabolomics experiment would most likely be conducted on a 600 MHz (14.04 T) magnet. A 600 MHz magnet provides sufficient resolution for metabolite identification and is more accessible than magnets with higher fields such as 800 MHz (18.72 T), 900 MHz (21.06 T) and 1 GHz (23.4 T) [129]. Higher field NMRs are a much more common commodity in biological research due to the much needed high resolution requirements of protein NMR. Unsurprisingly, metabolomics researchers with NMR access tend to benefit from magnetic fields of 600 MHz or above.

As explained earlier, NMR is inherently low on sensitivity, but improvements in this area are being made. In addition to high field magnets becoming more commonplace, the latest developments in cryogenically cooled probes, known as cryoprobes, have enhanced NMR sensitivity by at least an order of magnitude although it is still below the sensitivity of MS. Even with the latest cryoprobes and high field magnets it is not always possible to resolve a ^{1}H spectra. Usually, this problem can be overcome with adding information from a complementary nucleus such as carbon which is achieved by detecting another range of frequencies [120]. The carbon frequency range is broader than hydrogen (^{13}C range: 0-200 ppm) and so yields more resolved resonance peaks, although the NMR active isotope of carbon,^{13}C, is only abundant in nature at 1.11%. This results in longer data acquisition as only information from NMR active isotopes is able to be recorded [123].

The simplest NMR experiment is a single pulse for a single nucleus type prior to data acquisition, but they can be far more complex. A pulse sequence is a series of carefully designed and executed pulses to acquire specific types of information from the sample. Using different pulses *e.g.* different pulse configurations, power, and length of pulse, different types of information can be obtained from the sample [128]. For example, a sample containing both large and small molecules can be acquired using a 1D Nuclear Overhauser

Spectroscopy (NOESY) where all signals are observed. If the user requires to observe only the small or large molecules it is possible to do so without any chromatography. A Carr-Purcell-Meiboom-Gill (CPMG) experiment would suppress the signals arising from large molecules [128] whereas a longitudinal encode decode (LED) experiment will do the opposite and suppress small molecule resonances while large molecule signals remain intact [128]. One of the great advantages of NMR is that it is possible to observe multiple nuclei during a single experiment. The most common application of this is the use of 2D NMR experiments such as ^{1}H-^{13}C HSQC. Such experiments allow the deconvolution of signals utilising the chemical shifts on ^{13}C. This type of experiment observes the ^{1}H nuclei directly bonded to ^{13}C *i.e.* CH, CH_2 and CH_3 groups. HSQC experiments are a great aid in the identification of metabolites with high confidence. The ability to detect ^{13}C labelled materials in NMR also allows the probing of the fate of a site-specific labelled metabolite qualitatively and quantitatively in processes known as trace analysis [130] and metabolic flux analysis [131] respectively. The NMR experiments used for the work detailed in this thesis include 1D, NOESY, CPMG, and ^{13}C-^{1}H HSQC are covered in detail in section 2.3.4. Benefiting from high reproducibility, easier quantification and non-destructive nature NMR is a powerful tool to be used in metabolomics research.

1.6.5.2 Mass Spectrometry

Mass spectrometry (MS) is one of the most used analytical methods, with applications in compound detection and validation, purity assessment and protein sequencing, to name but a few [122]. As one of the two main analytical platforms (the other being NMR), MS dominates the field. MS is intrinsically highly sensitive in metabolite detection and quantitation as well as structure elucidation, within its limitation [Gowda & Djukovic]. MS is an analytical technique that involves interpretation of ionisation patterns of molecules in order to identify them [122]. A mass spectrometer is comprised of three essential components: an ion source, a mass analyser and a detector. These measure the mass over charge (m/z) of ions, with this information used to infer the identity of the metabolites, as well as their relative quantities. Typically, a sample is injected into the ioniser, which converts the metabolite into ions. These ions are then transferred to the mass analyser where ions are separated according to their mass over charge. Finally, a detector will capture the mass over charge and quantify the information before transferring it to a computer for analysis (Figure 1.6-5). The finer details of how MS works is beyond the scope of the work carried out in this thesis, but it is extensively covered in numerous publications and textbooks [122], [132], [133]. Instead, the use of MS in metabolomics will be discussed.

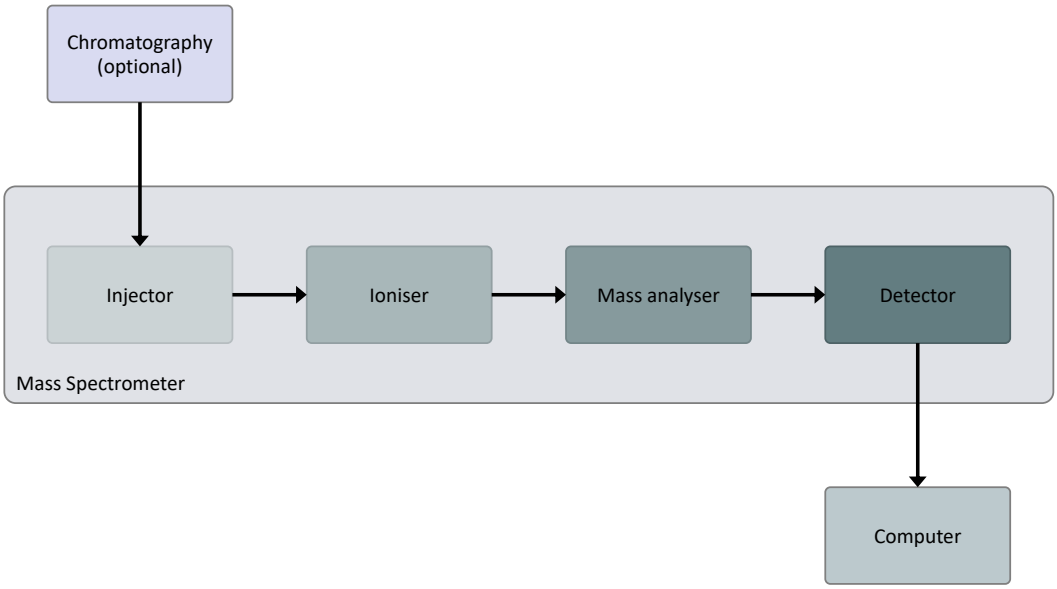


Figure 1.6-5: Schematic of a basic MS. Samples can either be injected directly or can be separated through a chromatographical method *a priori*. Injected samples containing the neutral sample are converted into ions *via* the ioniser. Molecular ions are then passed through the mass analyser which separates the ions according to their mass over charge. The detector records the mass over charge and relative abundance which is processed by a computer for analysis.

MS is a highly modular method and is often used in tandem with chromatography equipment (Figure 1.6-5) such as liquid chromatography (LC) or gas chromatography (GC). This functionality allows MS to be used in the analysis of biofluids, tissue samples and volatile compounds [134]. High performance LC-MS (HPLC-MS) requires samples to be in solution. HPLC separates the molecules present in the solutions according to their affinity to the attached column, examples of which include, C8, C18, C30 and hydrophilic interaction liquid chromatography (HILIC) columns. Once separated, these samples are individually sent to the MS for mass analysis. GC-MS requires samples and the compounds they contain to be volatile. In GC, samples are heated until they turn into gas where they can then be separated by their volatility, amongst other properties (e.g. size, polarity). Like LC-MS, once separated the compounds are sent in turn to a MS mass analyser. This technique is ideal for studying volatiles such as pheromones, although it can be used for a large range of metabolites [134]. The LC-MS approach is one of the most common in MS metabolomics due to its many benefits, but one of its biggest drawbacks is retention time drift. Using the retention time of molecules alongside mass over charge greatly increases the number of molecules that are able to be identified. However, this method is very sensitive to the experimental conditions. Changes in elution buffer composition, the age of the column and temperature can all cause

a change in the retention time of a molecule. This can happen between samples, batches or experiments. A drift happening in a single run or between batches of an experiment would greatly affect the structure of the dataset as well as the identification and may require a repetition of the experiments. Additionally, retention time drifts make comparison or replication of experiments from different labs a major challenge.

Ionisation is arguably the most critical stage in a MS metabolomics study. One of the limitations of MS is that only the detection of ionised molecules is possible. Understandably, the existence of a range of ionisation methods is not surprising. Electron impact (EI) and electrospray ionisation (ESI) are the most common ionisation methods used in MS metabolomics [122]. EI is a hard ionisation method where a heavy fragmentation of the molecule takes place and is the preferred method of GC-MS studies. In LC-MS studies, ESI is typically preferred. ESI is a soft ionisation method where less fragmentation of the molecule takes place. ESI is typically preferred due to is efficiency in ionising a wide range of small molecules, as well as larger molecules such as peptides and proteins. Another limitation of the ionisation specific to LC-MS is ion suppression. Ion suppression occurs when co-eluting molecules compete for the limited molecular ions. In such cases, molecules with low electron or proton affinity can be obscured or not detected [122].

In combination with its inherent sensitivity and ability to be easily coupled to chromatographic methods, MS is a great tool for qualitative analyses such as metabolic library building for organisms. On the other hand, a major limitation with MS is the challenging quantitative analysis. In quantitative complex mixture analysis, comparable quantitation of samples is of great importance. MS can only measure ionised molecules, and for a reliable exact quantification a non-variable ionisation is critical. Unfortunately, ionisation in MS is variable and so signal intensities do not always reflect the actual quantities in the samples. Another commonly referred to disadvantage of MS methods is the inability to identify isomers [135]. Since isomers are different configurations of the same molecule their masses cannot be used as an identification method, although it should be noted that it is possible to identify isomers with tandem methods such as HPLC-MS and/or MS-MS. Some isomers elute at different retention times and using their elution order, different isomers can be identified [136]. However, this method heavily relies on the chromatographic method and in some cases can even be performed without the mass spectrometer *via* different columns [136]. Another approach to identify isomers is to use MS-MS. Some isomers can exhibit characteristic fragmentation patterns which can be recorded with MS-MS and identified

[137]. Although possible, MS isomer identification is not a straightforward and robust method. It requires a lot of user interaction and may not be feasible for a large metabolomics study. Due to the nature of the MS systems, cross contamination between samples is a common problem, which brings further issues with reproducibility and quality assurance (QA). In order to monitor these problems, MS metabolomics experiments require QC samples in the data acquisition. Typically these are a series of pooled samples to ensure the quality and integrity of data acquisition [138]. Furthermore, this approach compounded with a hyphenated technology (e.g. HPLC) means the sample amount for a metabolomics experiment dramatically increases. Although, even with QC sample concentration required are very low.

1.6.5.3 Advantages and disadvantages of NMR and MS in metabolomics

Between the two leading methods in metabolomics, NMR and MS, the latter tends to be the preferred choice. Besides the numerous advantages of mass spectrometers; they are cheaper than NMR spectrometers, more space efficient (NMR requires anti-vibration measures on the floor for high quality data acquisition), and mass spectrometers can be operated with less training than is needed for NMR. Although these are not the only reasons why MS is generally preferred, they do make MS more widely accessible and hence more widely adopted which perhaps can be considered as an advantage. As a side-effect of being more widely used, MS has larger curated databases aiding in metabolite identification on a larger scale, especially with the added dimension of retention time (acquired from chromatography). This, compounded with the sensitivity down to nM (pM for some equipment) makes MS a very versatile instrument in metabolomics research.

LC-MS measurements performed on the same instrument in the same laboratory can present systemic variability in data collected over long periods of time [139]. These variabilities typically manifest themselves as changes in ionisation (co-elution of metabolites, salts, pH, mobile phase), drifting in retention time (degrading or replaced column), and mass calibration drift (temperature or circuitry configuration). Such variabilities can be observed over long acquisition runs or in the same set of samples acquired in 2 batches, although there are special measures that can be taken to deal with these issues. NMR on the other hand provides a framework which is more resistant to such systemic variabilities. Similar to MS, NMR can also be affected by temperature changes, pH variation, and salt concentration but in contrast they are easier to resolve. NMR samples sits on a temperature-controlled probe

which can be calibrated to 0.1 ˚C (see section 2.3.1) and can be monitored throughout the acquisition time. Variations in pH are typically seen as shifts in peaks, but can be stabilised using NMR compatible buffers. Although this increases the ionic strength of the solution, NMR pulses can be calibrated per sample for consistency of data acquisition. The ease and robustness of such a calibration makes NMR exceptionally reproducible and allows for acquiring data in batches as well as acquisition of additional experiments at a later date.

There are cases in metabolomics studies where re-acquiring of data or additional data is required. In such cases, using the same sample is the gold standard. Excluding the disadvantages of reproducibility of methods, another fundamental difference between NMR and MS can be seen. MS is a destructive method where the sample injected is not recoverable, whereas NMR samples are always contained in the sample tube and can be stored and reused. That being said, a critical point that needs to be mentioned is the amount of sample 'destroyed' in a MS injection is far less than the amount required to acquire an NMR spectrum. Nevertheless, NMR is still the only method where the initial sample remains untouched.

A common challenge for both methods is quantification. It is commonly referred to that NMR is quantitative (see section 1.6.5.1) and MS is not. This is a common oversimplification and is not entirely correct. It is possible to have quantitative measures, be it absolute or relative, with either of the techniques if proper methods are implemented. Nevertheless, this is much easier achieved with NMR than it is with MS. As explained in section 1.6.5.2, MS signal intensities can be varied due to concentration, efficiency of separation and ionisation conditions and so quantification requires careful calibration of the equipment, use of internal standards, and spiking of samples for accurate quantification. In contrast, for NMR, relative quantification is readily available on data acquisition and for absolute quantification the user can chose from: internal standard, external standard or a calibration curve.

Another key detail to be aware of is the polarity of the compounds of interest. Typically, metabolomics studies are conducted on either water soluble (polar) or non-water soluble (apolar) compounds. In both LC-MS and NMR, apolar compounds are typically more straight forward to measure. In NMR, samples are mixed with a suitable solvent and data acquired. For LC-MS, chromatographic methods are well established for apolar compounds, but separation of polar compounds is often more challenging. When more common place chromatography columns (e.g. C18) are used, it is often complicated due to the poor

interaction of biologically relevant polar compounds with the column resulting in fast elution of compounds and poor separation [140]. For such applications, the relatively new HILIC columns can be of great help. These columns have a higher affinity to polar compounds, eluting compounds in order of increasing polarity thus achieving a greater overall separation [140]. However, they are generally costly and retention mechanisms are not well understood [140]. In comparison, NMR acquisition of polar compounds requires only dissolution in an appropriate NMR buffer before acquisition.

Lastly, LC-MS platforms use a flow system for sample transportation meaning all samples are carried through the same tubing. Using one route for all the samples brings the possibility of cross-contamination of samples, prior to and during a run [138]. In comparison, NMR is a closed system where samples never come into contact. Although cross-contamination is a concern that the user should be aware of, with the use of proper pooled sample checks to track the status of consistency [138], and performing periodic maintenance of the equipment with cleaning and purging, cross-contamination can be kept in check.

Often in metabolomics, a researcher will have to choose between one of these techniques. Ideally, they would be used in combination to provide complementary information, but this is rather uncommon due to financial constraints, the expertise needed to operate both technologies in a metabolomics appropriate manner, instrument availability and sufficient sample/material. In a situation like this, the strengths and weaknesses (Table 1.6-1) of both techniques should be considered to choose the better suited technique for the question at hand.

Table 1.6-1: Strength and weaknesses of NMR and MS.

		NMR	MS
Sample	Recovery	Non-destructive analysis	Destructive analysis
	Volume	Higher volumes	Lower volumes
	Large molecule filtering	Non-destructive filtering	Sample processing required
Instrument	Quantitation	Absolute and relative	Relative
	Calibration and setup	Easier	Harder
	Installation cost	Expensive	Cheaper
	Maintenance & operational cost	Cheaper	Expensive
	Time per sample	1D: short time 2D: long time	Long data acquisition when coupled to chromatography
	Reproducibility	Reproducible	Semi-reproducible
	Sensitivity	Lower sensitivity	Higher sensitivity
	Dynamic range	Larger	Smaller
Data	Metabolite identification	Challenging	Easier
	Available databases	Limited	Several
	Size of databases	Smaller	Larger
	Analysis applications	Limited	Several

Acknowledging that MS would be a much more sensitive choice for identification, an in-house pilot test performed on single mosquitoes showed that certain NMR configurations (i.e. 700 MHz equipped with an inverse cryoprobe) are sensitive enough to acquire data to the complexity that allows the observation of experimental variations. Furthermore, considering the high reproducibility (e.g. use of external databases for identification), reusability of sample (i.e. for further experiments or repeats), ease of acquiring polar metabolite data with minimal steps, and eliminating the possibility of cross-contamination at the data acquisition stage, NMR was selected as the preferred analytical platform in this project.

1.6.5.4 Metabolomics datasets and their analyses

Once data is acquired we can proceed to the data analytic steps (Figure 1.6-6). Given that the metabolomics platform chosen for this project was NMR, in this section the data analysis steps undertaken when analysing ^{1}H NMR spectra will be described. It should be noted that majority of the downstream analyses are platform independent and is stated where applicable. Prior to any data analysis checkpoints to assess sample quality control (QC) need to be in place. These QC steps typically include markers for suitable acquisition and assessment of overall sample variance to identify potential outliers. The later should be

match with experimental notes to ID any problems in current pipelines and iteratively improve the sample acquisition steps.

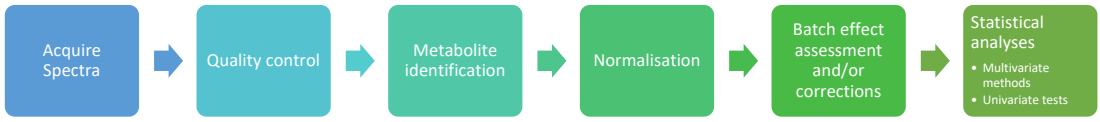


Figure 1.6-6: Diagram representing the analytical steps taken in classic metabolomics analysis.

Sample failing QC due to faulty acquisition may be acquired again with the aim of retaining as many samples within the overall study. Upon consistent failure (typically 3 tries) failed samples should be identified as problematic and removed from the study. Typically NMR QC, includes the assessment of the half height full width of the internal standard, overall shape of the baseline, wideness of the residual water peak and signal overflow. NMR QC is explained in detail in section 2.3.6.

Generally, data acquisition is followed by metabolite identification. This is typically done by a combination of in-house metabolite library and/or external databases (e.g. Chenomx (Chenomx, CA), HMDB [119], BMRB [141]) where the spectra are compared to spectra of standard compounds. Details of the metabolite identification is detailed in section 2.4. Typically, in NMR dataset there are two choices either absolutely quantified datasets (will be referred to as absolute dataset) or relatively quantified dataset (will be referred to as relative dataset). Absolute datasets contain concentration values per metabolite per sample. As identification is often incomplete, these datasets contain fewer variables but are more powerful as results from their analysis are easier to interpret biologically. In these datasets, unidentified signals are removed. On the other hand, datasets of relative quantification (relative datasets) contain all the information acquired by all the signals independently on whether they are identified or not. In NMR a metabolite may have more than one signal and that signal might have an overlap with others, what can make interpretation more difficult. However, these datasets retain most information and can be key to detect relevant biological samples. It is important to mention that a step to assess the quality of the signals of each of the metabolites identified needs to be in place in order to select the best suitable marker for that metabolite. There are currently no standard methods to do so but I propose one in section 2.5.2.3.2. Best operating practice include the provision of relative datasets within any publication and accompany them from the absolute dataset if relevant. This allows for a in-

depth comparison of both and allows for further analyses by the community that might have access to wider metabolite libraries to further identify signals.

Metabolomics experiments typically require the acquisition of large number of samples which typically cannot be processed within the same experiment. This implies the need for sample acquisition in batches. Variation that may be introduced from sample handling, pipetting errors, batch and/or data acquisition needs to be controlled. Such variations may mask the experimental variation of interest. The first step for unwanted variation control is normalisation. There are multiple normalisation methods used in NMR metabolomics, some of the most common include normalisation by total area, normalisation by a reference spectral bin, and probabilistic quotient normalisation (PQN). Total area normalisation is most commonly used when the suspected source of non-experimental variation is mainly due to dilution effects [PQN ref]. Normalisation by a reference spectral bin is typically preferred when a given standard has been added to the metabolite mixture prior to extraction and can serve as representative of the proportional changes a sample has suffered in the extraction and acquisition steps. It can also be a typically stable signal in particular biofluids, for example creatinine is widely used to normalise 1H-NMR urine spectra. PQN normalisation method can be used for both relative and absolute concentration datasets. This normalisation method is designed to calculate a normalisation factor through the quotients of median values in the dataset (see section 2.5.1.2 for details). Due to the use of medians instead of means in PQN, it is inherently resistant to the skew caused by outliers and sample with extreme cases (which typically may be confused with outliers). in the spectra an is applicable to both relative and absolute concentration datasets [PQN ref].

Normalisation methods are not always sufficient due to the high batch effects in the dataset. In such situations additional batch correction methods also need to be applied. Methods for batch correction develop for omics data analyses (i.e. multidimensional data) are based on analysis of covariance (ARSyN [142]), Bayesian models (ComBat [143]) or Surrogate Value Analysis (SVA [144]). Detailed explanations of these methods would be out of the scope of this project hence the reader is encouraged to read the detailed publications of methods themselves. All batch correction methods mentioned above use different statistical approaches to minimise the non-experimental variation. Some of these methods are more powerful than the others and all of them may introduce some bias to the dataset. For example, ARSyN is designed to remove any variation that is not explained by the experimental groups provided (for example drug treatments). This typically yields excellent

results separating the groups of interest but it can introduce bias by removing variation that might be also relevant to it (for example sample demographics such as sex or age). On the other hand, ComBat is a method designed to remove the variation linked solely to known batches. This might lead to worse outcomes but it is a safer option if we want to keep other possibly relevant variation within the data. Suitability of batch correction methods should be assessed to choose the most suitable method. Principal variance component analysis (PVCA) [145] is an algorithm that estimates the variation in the data based on the factors provided such as, experimental factors, non-experimental factors (e.g. batch), and residual variation (i.e. all the remaining variation in the dataset). PVCA can provide information of the variation reduced by a particular batch effect correction method.

NMR signals represent specific nuclei in the molecule under specific magnetic environment they experience. Thus, a metabolite's NMR signature can be a single peak or more (see section 1.6.5.1 more detail). Thus, NMR data is highly multicollinear.

As metabolomics datasets are multivariate, it is advisable to use multivariate methods for their analyses. These can be unsupervised or supervised. Unsupervised methods does not require the experimental-group information. Most widely-used multivariate transformation in omics analysis is principal component analysis (PCA) (see section 2.5.2.1).

PCA is typically used as a data exploration tool where the inner structures of the data can be revealed. PCAs are very fast to calculate and are implemented in many statistical packages. When a structure cannot be appraised with PCA, supervised methods can be used. These methods use the information regarding the groups present in the data to maximise differences between groups. Examples to such methods are partial least square discriminant analysis (PLS-DA) or random forest (RF). These methods and their model building procedures are covered in detail in section 2.5.2.2. these methods build a model that uses a fingerprint of signals to predict the different groups. The importance of each of those signals is rated, thus a selection of most important metabolite signals can be subsetted and taken forward for more analysis. Supervised methods create models. These models are prone to overfitting. Overfitting is the process by which a model performs excellently to describe the data used to build it but is unable to predict groups from new datasets (i.e. it has fitted to noise). In order to avoid overfitting rigours cross-validation steps are required and are covered in section 2.5.2.2. PLS-DA models are widely used in metabolomics and implemented in most analytical

steps. There are not the best performing supervised algorithms but they are most appropriate for multi-collinear datasets such as NMR.

Once a multivariate approach has led a list of interesting features/metabolites, one can further the analysis by undertaking 2-sample or 2+ sample tests such as Student's t-test or ANOVA test respectively. These hypothesis testing methods are evaluated by using an arbitrarily set threshold known as the α value and is typically set to 0.05 (or 5%). This threshold represents the amount of potential false positives we are willing to accept in our comparisons. As the names suggest univariate tests are designed to compare a single variable and are not designed for multiple hypothesis testing as it is. Meaning testing of multiple simultaneous variables from the same data. When a univariate test is used in multiple comparisons, it becomes increasingly more likely for at least one comparison to be different with statistical significance (determined by the α = 0.05 threshold). Thus, by using multiple comparisons the probability of making an error increases with each test made. Consequently, when undertaking multiple hypothesis testing we need to incorporate a step to correct for potential false discoveries, to avoid falsely concluding some metabolites to be important while they are not. Two methods can be commonly observed in metabolomics as well as biological sciences in general to undertake false discovery corrections. The more conservative method Bonferroni correction and the less conservative Benjamini-Hochberg method. In Bonferroni method randomly assigns alpha-new values (can be variable) to each test such that the sum of all alpha-new values is equal to the desired α value for the entire test (typically 0.05). This method is very conservative in controlling false positive discoveries, but the trade comes with cost of increased discovery of false negatives. The alternative method is the Benjamini-Hochberg (BH) method. In this method p-value are ranked in an increasing fashion and a new alpha threshold is calculated per comparison via $\left(\frac{i}{m}\right) Q$ where i is the raw p-value, m is total number of comparisons, and Q is the alpha value for the test. Then each raw p-value is compared agains its new threshold. This step is then followed by finding the highest raw p-value that is lesser than its new α threshold. Any comparison before this is considered significant and this process is continued until all comparisons are evaluated. This method controls the false discovery rate as well as the false negative rate thus, have a higher statistical power compared to the Bonferroni method. Note that univariate methods can be applied to multivariate tests but their use is not efficient providing the false discovery rate increases with the number of tests. Thus, although the option is available within multiple

statistical packages (including Metaboanalyst [146]), it is advised to undertake a multivariate approach followed by a univariate one.

Once a robust list of relevant variables whether identify with multivariate methods or jointly verified with univariate methods, the next step in the pipeline is the biological contextualisation. In metabolomics this is typically combined with pathway analysis. Pathway analysis methods are one of the bottlenecks in metabolomics analyses. Most common implementations are based on Fisher's exact test (see section 2.5.4 for details) where the probability of a set of input metabolites representing another set of metabolites (i.e. pathways) is calculated. Such methods are typically called metabolite set enrichment analysis (MSEA). Users must be cautious when using MSEA methods since the results of such test are not definite answers and should be treated as leads to follow up un. It should be noted that MSEA methods do not take metabolite levels into account. Additionally, the quality of such tests heavily depends on the quality of pathway database used. It should be noted that there are implementation of MSEA where pathway rankings can be calculated through an 'importance' score. These implementations typically depend on the number of connections the metabolites have in the pathway (usually calculated via in-betweenness centrality). Although it may be more informative, similar to MSEA the result of such analyses should be taken as a complementary metric to the results rather than definite answers.

In this project a relative datasets were used to incorporate as much as information from the data acquired. Normalisation of choice was PQN due its high effectiveness on NMR data and Batch effects were assess via PVCA and multiple correction methods were tested and applied per dataset. Multivariate analyses PCA and PLS-DA were used to explore data and to select variables respectively. Selected metabolites were compared using the univariate methods t-test and ANOVA as dictated by the number of experimental groups. Multiple comparison adjustments were performed using BH to have a more balance false positive and false negative correction. MSEA was performed by a custom implementation of Fisher's exact test and the data was contextualised by correlating with the information available on publications. Implementations of the statistical methods are covered in detail in section 2.5.

1.6.6 Metabolomics studies on insects

Since the turn of the century there has been a total of 22,856 documents published under the topic 'metabolomics' (data from Web of Science, Clarivate Analytics; UK) with 17,488

(72.82%) novel research articles (Figure 1.6-7). Within the 22,856 documents, 197 (0.86%) were published under the topics of metabolomics and insects but, only 25 (0.11%) of these were published under the topics metabolomics and mosquitoes, all of which were published between 2004 and 2019.

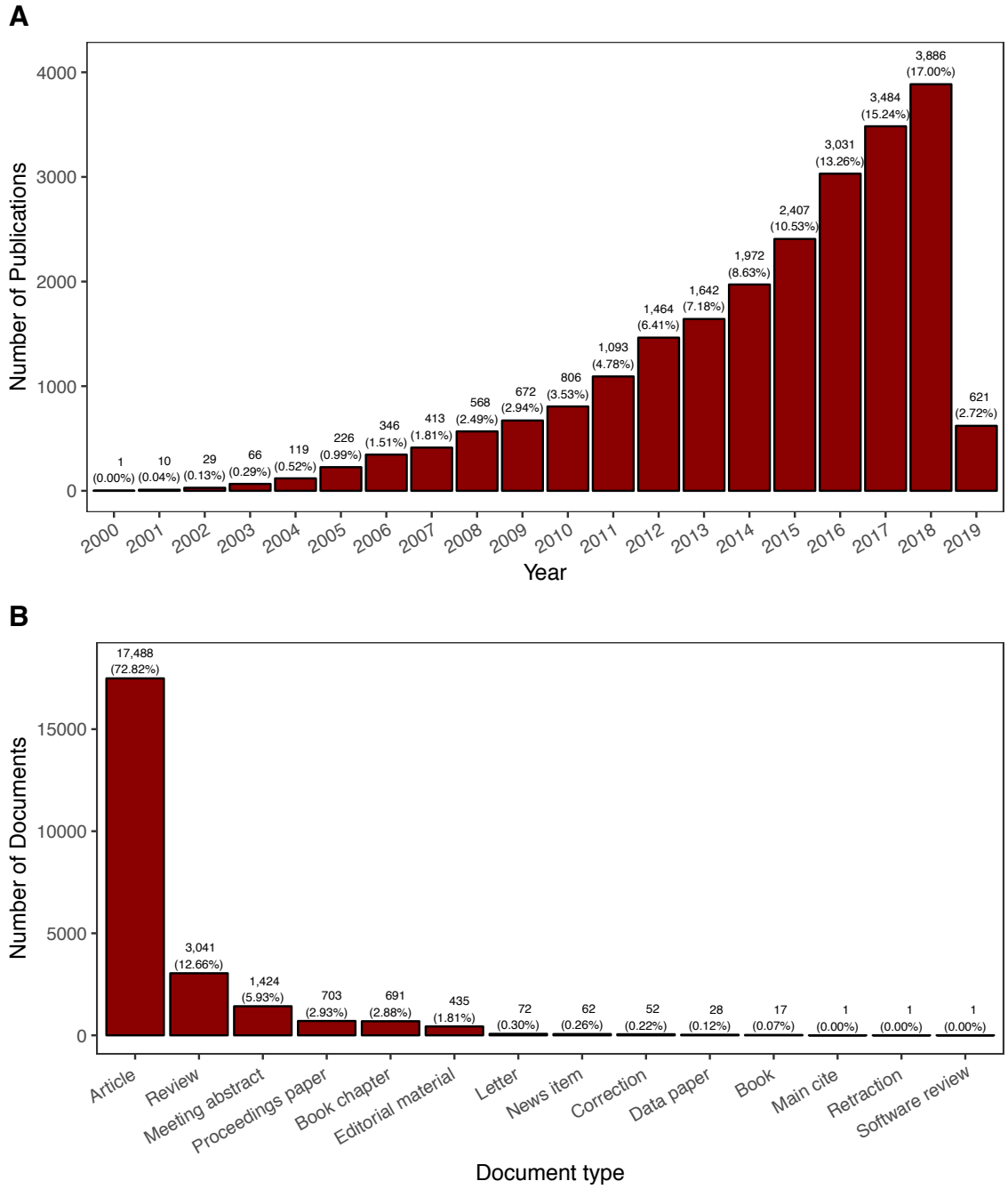


Figure 1.6-7: 'Metabolomics' publications between 2000-2019 data from Web of Science. A) Number of publications under the topic 'metabolomics'. Percentages are of the total number of metabolomics publications (22,856). B) Breakdown of the published documents topic 'metabolomics'. Due to some documents having multiple document type tags total number of documents are 24,016.

In 2015, Snart *et al*. published a mini review mentioning the term 'entometabolomics' (entomology + metabolomics). However, this term has not been widely used since. Currently, metabolomics is employed in insect research to investigate insect-bacterial interactions [147]–[149], insect-plant interactions [150], [151], development [152], interactions with commercial pesticides [153], [154] and temperature stress [155], [156]. The most published model to date is *D. melanogaster*. This is perhaps unsurprising considering the amount of scientific knowledge gathered using *D. melanogaster* [157], [158]. In terms of metabolic profiling, there is a keen interest in the profiling of insect biofluids, specifically the haemolymph. One of the most important milestones in insect metabolomics is the profiling of the insect haemolymph in 2008 (Phalaraksh *et al,* 2008).

Amongst the 245 unique publications under the topics metabolomics and mosquitoes/insects, two studied *Anopheles* mosquitoes and seven investigated *Aedes* mosquitoes (Table 1.6-2). Although a metabolomics approach is adopted by many fields of research, it is still not common in mosquito research.

Table 1.6-2: Research articles between 1900-2019 on Web of Science under topics: metabolomics and mosquitoes/insects which studied either *Anopheles* or *Aedes* species. CHIKV: Chikungunya virus, DENGV: Dengue virus, Ref: reference.

Species	Sample	Technique	Short summary	Ref
Ae. aegypti	Abdomen extract	LC-EI-MS	Crowding and starvation during breeding causes lower levels of alanine due to high turnover to pyruvate.	[159]
	Haemolymph	NMR	L-cysteine levels in the haemolymph are affected upon CHIKV and DENGV infection.	[160]
	Whole extract	GC-MS	Carbohydrate metabolism is controlled by methoprene-tolerant receptor, after a blood meal it is controlled by ecdysone receptor.	[161]
			Ibuprofen exposure increased male mosquito life span and in progeny eggs metabolic resource internalisation is affected.	[162]
		HPLC-MS	Metabolomics of juvenile hormones synthesis pathway and mevalonate pathway showed an inverse relationship.	[163]
		LR-LC-MS-MS	Glucose is used in ammonia detoxification after a blood meal.	[164]
Ae. albopictus	Whole extract	UHPLC-TOF	Nine lipids were identified as more abundant in quiescence eggs and tyrosine metabolism identified as different between quiescence and non-quiescence eggs.	[165]
An. coluzzii	Whole extract	GC-MS	Higher levels of triglycerides in populations more prone to desiccation	[166]
An. gambiae	Whole extract	UPLC-MS/MS	Revealed a parasitic relationship between the host mosquito and *Mycobacterium ulcerans*.	[149]

Amongst the studies published under the topics of 'metabolomics' and 'insects', desiccation resistance was studied by two groups. In 2018, Batz and Armbruster investigated the HC

composition of *Ae. albopictus* eggs, where they compared the changes when the eggs undergo quiescence [165]. In 2016, Hidalgo et al focused on the physiological adaptations to desiccation in distinct populations of female *An. coluzzi* and highlighted the mosquito response to desiccation through the use of metabolic processes [166]. Batz and Armbuster used a multiomics approach by incorporating metabolomics information into the transcriptomics data acquired from a previous studies. Their analytical platform of choice was GC-MS for lipidomics and LC-MS for polar metabolomics. Metabolite identification for lipidomics was performed by using an external library, whereas polar metabolites were identified *via* their in-house library. Their statistical analyses showed an accumulation of nine lipids and nine glycerides (two diglycerides and seven tri glycerides) in eggs that undergo quiescence [165]. One of the interesting results of their study was the identification of 36 distinct metabolite that were uniquely found in quiescence eggs, but unfortunately none of these metabolites could be identified [165]. The study undertaken by Hidalgo et al in 2016 investigated the desiccation in *An. coluzzi* [166]. Although the species of these two studies are different, and Anopheles eggs do not undergo quiescence, Aedes egg that undergo quiescence are resistant to desiccation [ref mosq book]. Aedes eggs can survive in considerably drier conditions and so it is not unreasonable to expect some of the desiccation resistance mechanisms to be similar. Hidalgo et al identified 36 metabolites, of which 29 showed significant differences between experimental groups [166]. These metabolites were mainly consisted of saccharides and amino acids [166]. When the two studies are compared, it is evident that polar metabolic profiles of adult mosquitoes and eggs are quite dissimilar. Although it can be speculated that some similar metabolites are either unidentified or extremely low concentration of egg metabolites were not quantifiable. Nevertheless, both studies have identified triglycerides as being important. Hidalgo et al argued the oxidation of triglycerides can provide a reliable source of water, whereas Batz and Armbuster discussed triglycerides as a reliable energy source. It has been reported that triglycerides have a high calorie content and release water on oxidation [#3 in Hidalgo refs]. Triglycerides are synthesised by the esterifcation of a glycerol with 3 hydrocarbons. An increase in HC production under desiccation conditions is therefore very likely in order that they can be used to make triglycerides which can be oxidised to produce the water which is lacking.

Another area of interest is the factors effecting the development of mosquitoes. A 2015 study by Price et al investigated the *Ae. aegypti* development using a combination of metabolomics and transcriptomics. In this study, one strain was bred under typical laboratory conditions, referred to as large mosquitoes, and another was raised under crowded and nutritionally

deprived condition, referred to as small mosquitoes [159]. All samples used in this experiment had the fat body obtained through dissection of the mosquitoes. Samples were then pooled to create samples for GC-MS and RNA-sequencing [159]. The obtained transcriptomics results indicated the vitellogenesis (egg yolk formation) process to be negatively affected in small mosquitoes. The transcript counts were not statistically different prior to a blood meal, although this was expected due to vitellogenesis being activated only after mating and a blood meal. In large mosquitoes there was a 1.6-2 fold increase in vitellogenesis transcripts [159]. The researchers interpreted these results as being due to the fact that in small mosquitoes all the nutrients obtained from the blood feeding were redirected to energy related mechanisms in order to compensate for the malnourished state they were in. As complementary information, Price et al also used metabolomics. An inconsistency observed in this publication was the vague description of the analytical platform used for the metabolomics. The metabolomics section mentions LC-MS, although the methods section describes a derivatisation procedure which is commonly used in GC-MS. Since the analytical platform is not explicitly detailed, extrapolating from the mentioning of a derivatisation procedure it is assumed that GC-MS was used. Price et al identified 138 metabolites of which 58 were different between small and large mosquitoes after a blood meal and 65 were different for no blood meal [159]. They showed a particular interest in the amino acids and the decrease they observed between small and large mosquitoes, interpreting the results as the amino acids being used as energy sources rather than resources for protein production. This study revealed some interesting differences between different breeding conditions, although the experimental design includes multiple variables which may not be deconvoluted. More specifically, between the large and small mosquitoes there is the variation of nutrition provided as well as crowding of the breeding tray. The majority of their conclusions are related to energy metabolism, but it is not possible to determine if this is related to the crowding, nutrition, or perhaps a mixture of both. A follow-up study could possibly shed more light on the metabolomics of breeding of mosquitoes.

In the same year as the Price study, Hou et al investigated the reproductive cycles of mosquitoes in relation to the juvenile hormone and its role in carbon metabolism in *Ae. aegypti* females using transcriptomics and metabolomics [161]. Hou et al used GC-MS metabolomics in addition to microarray transcriptomics for a multiomics approach. Transcripts of carbon metabolism after eclosion and post blood meal were compared as a time course study. These transcripts were found to be significantly higher in the first 24 hours after eclosion, decreasing afterwards until a blood meal [161]. After a blood meal these

transcripts were again found to be significantly elevated, followed by a decline over a period of 48 hours. In order to verify their results, they used GC-MS metabolomics to quantify the intermediary metabolites of the elevated carbon metabolism enzymes (proxied through transcripts). Between the two sources of data, a high correlation was observed in enzymes and their final product. An interesting outcome of their research was the distinct citrate cycle profiles between post eclosion and post blood meal [161]. Considering that in both cases carbon metabolism is upregulated, it is reasonable to expect the citrate cycle profiles to be similar. However, in their findings, citrate and maleate were found to be significantly reduced in the post eclosion phase, whereas they were found to be higher after a blood meal [161]. The juvenile hormone and its receptor, methoprene-tolerant receptor, are both key to the regulation of a mosquitoes reproductive cycle [161]. In order to test their role in carbon metabolism, the juvenile hormone and receptor were applied topically on newly eclosed females. The application caused a significant decrease in carbon metabolism transcripts and an elevation in glycogen and glucose compare to untreated mosquitoes [161]. Thus, indicating Methoprene-tolerant receptor controlling the carbohydrate metabolism pathways upon a blood meal. As revealed by Price et al, in malnourished conditions a blood meal can be redirected to energy metabolism rather than vitellogenesis. This 'switch' was also investigated by Horvath et al in 2018 by analysing *Ae. aegypti* fed with blood meals supplemented with ^{13}C-glucose using LC-MS [164]. Mosquitoes were fed with a 3% sucrose solution for 3 days then were subjected to a 24 hour starvation period, followed by the ^{13}C glucose supplemented blood meal [164]. The starvation step is necessary to remove the interference of diet, but as Price et al showed, starvation applies a pressure for the blood meal to be utilised in energy production which was not mentioned in the study [164]. In this time course study, Horvath et al traced the ^{13}C revealing the utilisation of glucose in ammonia detoxification and uric acid synthesis upon blood feeding through the glycolysis and pentose phosphate pathway [164]. This study made a note that although the acetyl-CoA derived from the ^{13}C glucose is expected to be used in fatty acid production based on literature, the measuring of fatty-acids was out of the scope of the study. Similarly, it would be reasonable to expect a positive correlation between energy mechanisms and fatty acid production.

The ^{13}C glucose trace study by Price et al revealed the role of glycolysis in ammonia detoxification [159]. Ammonia is a by-product of the processing of the blood meal, and it is not the only dangerous one. Processing of a blood meal creates a great oxidative stress on the mosquito. In 2018, Shrinet et al investigated the effect of this oxidative stress and how it is amplified when the blood is infected with arboviruses such as chikungunya virus or dengue

virus in *Ae. aegypti* using a multiomics approach with the integration of transcriptomics (published data), proteomics (unpublished data) and metabolomics (experimentally acquired NMR data) [159]. To understand the effects of the oxidative stress under arbovirus infection, mosquitoes infected *via* injections with chikungunya virus, dengue virus, and co-infections of these viruses were compared. NMR was performed on a 500 MHz magnet fitted with a cryoprobe for high sensitivity. In contrast to most studies, Shrinet et al used the haemolyph of the mosquitoes rather than whole extracts in this study [159]. Shrinet et al did not report the total number of identified metabolites, but their PLS-DA models yielded 15 metabolites of importance *via* VIP scoring [159]. Following the PLS-DA modelling, a pathway analysis was performed. Taurine and hypotaurin pathways were the topmost predicted pathways followed by pantothenate and CoA biosynthesis, however, the list of metabolites used for the pathway analysis was not reported [159]. Following the metabolomics analysis, an integration of proteomics and transcriptomics data was performed. Firstly, published transcriptomics data was selected from the literature and supplemented by an unpublished in-house proteomics dataset. These data were also analysed with pathway analysis individually and the results were merged in the discussion. A common pathway highlighted by all the methods was oxidative phosphorylation [159]. This was attributed to the virus modulating the bioenergetics of the mitochondria in order to optimise the viral replication [159]. Furthermore, L-cysteine levels were found to be lower in infected mosquitoes affecting the taurine and hypotaurine pathway during oxidative stress. L-cysteine is a precursor of glutathione which is an antioxidant and was not reported by this study [159]. Nevertheless, the study concluded that oxidative stress caused by the blood meal can be further increased by an arboviral infection on which upon *Ae. aedes* species can inhibit oxidase in order to stop the production of reactive oxidative species.

Viral infection in mosquitoes does not only create oxidative stress, but can also affect the development of the mosquitoes. The 2015 study by Hoxmeier et al investigated *An. gambiae* mosquitoes infected with live and dead *Mycobacterium ulcerans* [149]. *An. gambiae* larvae were reared in water with live *M. ulcerans*, dead M. ulcerans (gamma irradiated), and no *M. ulcerans* [149]. Emerging adult mosquitoes were collected for metabolic profiling using LC-MS. Upon analysis, all three groups presented distinct metabolic profiles [149]. Most notably, the phospholipid pathway was affected when exposed to live *M.* ulcerans. Di-acylphospholipids were higher in the live *M. ulcerans* infected mosquitoes as well as tri-glycerides [149]. Amongst the findings were also lower survival and pupation rates during development [149]. In light of the findings, a parasitic relationship between the bacteria and

the host rather than a communalistic relationship is implied. Unfortunately, the depth of disruption to the phospholipids and triglycerides was not covered in this study and it is a relatively unexplored area. It may be of great importance to know whether these disruptions go as far as altering the hydrocarbon productions as such alterations might interfere with the delivery of insecticides (i.e. cuticular hydrocarbon resistance).

The use of metabolomics in mosquito development is relatively more common than its use in other mosquito research areas. A 2018 study by Prud'homme et al investigated the direct exposure of ibuprofen on *Ae. aegypti* adults and their progenies using a combination of metabolomics and transcriptomics [150. In their study, Prud'homme et al observed no phenotypic change in the parents (F0). Although the survival rate of the ibuprofen exposed parents' first-generation progeny (F1), was twice as much as control under starved conditions [150. They have identified 134 differentially transcripted genes in the F0 larvae, whereas this number was 1566 in F1 larvae [150. The F0 larvae did not show any stress related transcripts [150. The metabolomics data was acquired *via* GC-MS on F0 males and females and F1 eggs (1-4 hours old). The F0 metabolic profiles did not show any differences between control samples and those exposed to ibuprofen [150. In contrast, F1 eggs showed 20 metabolites to be significantly lower in the ibuprofen exposed group [150. These metabolites mainly consisted of amino acids, carbohydrates, and polyols, thus indicating impaired resource internalisation during embryonic development [150. Interestingly, upon hatching these exposed F1 larvae showed overactivation of the ecdysone signalling pathway which promotes accelerated development [150. This over activation was discussed to be the reason for the F1 progenies high survival rate during starvation [150. It is unclear whether ibuprofen exposure is the reason for the high survival rate in the F1 progenies or this is simply a stress response. Nevertheless, it is clear, as shown by Prud'homme, that such pollutions in water bodies may cause more favourable survival conditions for mosquitoes, hence increasing the risk of vector borne diseases.

Lastly, Rivera-Perez et al studied mosquito development in 2014 with a focus on reproductive physiology in *Ae. aegypti*, particularly, mevalonate and juvenile hormone synthesis pathways and their links to the reproductive system [163]. The juvenile hormone is synthesised through the mevalonate pathway [163]. Being a multipurpose pathway, it is responsible for the biosynthesis of essential signalling molecules used in energy homeostasis and glycosylation [163]. Through the use of acetyl-CoA, this pathway branches into juvenile hormone synthesis [163]. Using chromatography coupled with a fluorescent detector and MS, a targeted

metabolomics approach was utilised. Quantification was done by adding fluorescent tags to the metabolites of interest, such as mevalonate and acetyl-CoA, and acquiring the data by fluorescens detector. In order to complement the metabolomics data, published studies were utilised for a meta-analysis [163]. In their results, they reported a 10-fold increase in juvenile hormone within the first 24 hours of eclosion [163]. Given its established role in development, it is expected to observe an increase of this hormone. Interestingly, the mevalonate pathway metabolites were low while juvenile hormone was on a high indicating a possible feedback loop between the two processes.

From the published literature covered in this section, it is clear that the majority of metabolomics studies on mosquitoes are in the form of a multiomics study, predominantly with transcriptomics. Furthermore, in metabolomics, the preferred analytical platform is GC-MS followed by LC-MS followed by NMR. It is also interesting that the majority of the discussions are derived from the transcriptomics data and supplemented by metabolomics, suggesting an underuse of the available metabolomics data. The metabolomics analysis techniques in the above literature are mostly univariate comparisons, thus the strength of multivariate analyses is rarely exploited. Excluding the studies on blood feeding, the majority of the studies only focused on female mosquitoes and the utilisation of male mosquitoes in research is severely lacking. Furthermore, a metabolomics study between males and females was not found. Given the strength of metabolomics to compare profiles and identify differences, it is surprising that no study comparing insecticide resistant and susceptible strains was found. Interestingly, all of the studies covered above were either focused or linked to the development of mosquitoes. It is also clear that there is a lack of NMR based studies. Judging by the experimental designs of pooling samples for both MS and NMR studies, researchers might be under estimating the sensitivities of both analytical methods. Thus, a gap in single mosquito metabolomics is present and exploring the detection limits of these analytical platforms may open up opportunities in mosquito research.

1.6.7 Metabolomics studies on cuticular hydrocarbons

In the past five years, the importance of CHC in insecticide resistance has been gaining more attention in the entomology world. The majority of the previous research on CHC was conducted between 1970-1990 by Michael Locke [167]–[172]. More recent work in the field was led by the research group of Gary Blomquist [78], [88], [173]–[175]. They established that insect CHC is well conserved across insect species as well as identifying similarities of

their biosynthesis mechanisms. However, to date, a full characterisation of this process still does not exist. Although there are no research papers published on this subject under the topic metabolomics, between 1997 and 2018 there were 11 research articles and one review (data from Web of Science) using methods that are employed by the metabolomics studies (Table 1.6-3).

Table 1.6-3: Publication search on Web of Science with the topics cuticular hydrocarbon, resistance and mosquitoes between years 1900 and 2019. Ref: reference

Species	Type	Short summary	Ref
Ae. aegypti	Research	Revealed CHC of older age mosquitoes attract same sex while three days old mosquitoes are more attractive to opposite sex.	[176]
		Age determination of adult mosquitoes using composition of CHC compositions from head and thorax.	[177]
Ae. albopictus	Research	Desiccation resistance in mosquito eggs are linked to HC composition on the egg shell surface.	[178]
An. coluzzii	Research	CHC abundance alone does not provide desiccation tolerance.	[179]
An. gambiae	Research	Cyp4g16 functions as a decarbonylase in CHC biosynthesis.	[75]
		CHC composition is predetermined before adulthood and are altered before and after mating during adulthood.	[180]
An. gambiae & *An. coluzzii*	Research	Revealed simulated dry and rainy seasonal conditions does not effect CHC composition.	[181]
		2La chromosomal inversion does not effect CHC composition and increases non-branched alkane abundance.	[182]
An. stephensi	Research	Modelled resistance and genetic background through CHC profiles.	[84]
D. melanogaster	Research	Decreased CHC levels in Cyp4g1 knock-down strains and increased mortality.	[183]
		Multi resistance mechanisms including thicker CHC layer, xenobiotic metabolism and toxin excretion.	[184]
	Review	DDT resistance through altered cuticular layer attributed to Cyp4g1.	[185]

Literature searches on cuticular hydrocarbons and metabolomics did not yield a vast repertoire of publications. Of the papers found, most were either about insecticide resistance in relation to CH resistance or CHC composition and profiles. On the whole, most of these studies were in agreement with each other's finding.

A recent review by Kim *et al* (2018) took the RNAi approaches in insecticide resistance studies and focused on *D. melanogaster* 91-R strain (DDT resistant) [185]. Although not a study carried out on mosquitoes, this well understood model is a close enough relative to mosquitoes. Many mosquito studies are based on extrapolations from *D. melanogaster* studies to other mosquito species. In their review, Kim *et al* drew attention to the upregulation of Cyp6g1, Cyp6a2 and Cyp12d1 [185]. These CYPs are known to be involved in DDT metabolism although they are not the sole reason why this strain is resistant to DDT [185]. Additionally, Kim *et al* elaborated on the insecticide penetration. As discussed in section 1.3.3.4, this phenomenon, also known as cuticular resistance, is the result of

alterations in the cuticular layer causing reduced permeability to insecticides. The reviewers acknowledge that this mechanism is still largely unknown although they draw attention to a particular CYP, Cyp4g1 [185]. This enzyme catalyses a decarbonylation step in the long chain fatty acid biosynthesis and is critical for the production of CHCs [87], [183], [186].

A 2015 study by Gellatly *et al* investigated the knockdown effects of Cyp4g1 in *D. melanogaster* 91-R strain [183]. Deriving the Cyp4g1 target through literature research, Gellatly *et al* employed a UAS-RNAi knockdown model to observe the effects of Cyp4g1. The knock-down flies were studied by mortality bioassays and CHC analysis. From the mortality assays, the researchers observed a 25% increase in susceptibility to DDT on the resistant strains [183]. Furthermore, using a targeted GC-MS approach, four out of five HCs were found to be significantly lower in the knock-down strains [183]. Unfortunately, identification of the hydrocarbons was not given, although they were chosen from another study conducted by Strycharz *et al* [184]. A study on a similar premise was conducted in 2013 by Strycharz *et al* where ^{14}C labelled DDT penetration was investigated in *D. melanogaster* 91-R [184]. The control flies (Canton-S strain) used showed higher contact penetration percentage compared to the *D. melanogaster* strain [184]. To supplement these findings, both 91-R and Canton-S strains were investigated with transmission electron microscopy, revealing the 91-R strain possessed a thicker cuticle [184]. Furthermore, the CHC composition was analysed *via* GC-MS and no qualitative differences were found between the two strains although five HCs (9-triacosane, triacosane, pentacosane, heptacosdiene and heptacosane) were found to be significantly different between the two strains [184].

In 2016, Balabanidou *et al* conducted a study 'translating' the Cyp4g1 decarbonylase in *D. melanogaster* into *An. gambiae*. Cross-referencing the over-expressed CYPs of resistant *An. gambiae* with potential *An. gambiae* orthologues of Cyp4g1, two enzymes were identified: Cyp4g16 and Cyp4g17 [75]. These enzymes were also shown to be upregulated in insecticide resistant species. A deltamethrin uptake experiment was performed with resistant (Tiassale strain) and susceptible (N'Gusso strain) strains of *An. gambiae* where resistant strains were observed to internalise deltamethrin approximately 50% slower [75]. Furthermore, when cross-sections of these strains femurs were compared under a transmission electron microscope, the resistant strains were found to be significantly thicker, similar to the *D. melanogaster* 91-R strain [75]. Overall, CHC profiles of both strains were analysed *via* GC-MS and no qualitative differences were found [75]. Furthermore, HC levels were compared as a total with no individual HC levels being reported. In order to further characterise these two

enzymes, an immunostaining assay was performed, revealing both enzymes reside within the oenocytes, but are sub-localised differently. Cyp4g16 was found to be associated with the plasma membrane on the intracellular side whereas Cyp4g17 is dispersed throughout the cytoplasm [75]. Balabanidou *et al* tried to express both of these enzymes in order to evidence their role in decarbonylation. Unfortunately, only Cyp4g16 could be expressed as a fusion protein with CYP reductase at levels high enough to test its decarbonylase activity *in vitro* [75]. Balanbanidou *et al* showed evidence for cuticular resistance in *An. gambiae* and identified an enzyme which can be used to further study the mechanism that can give an insight into the function of CHCs.

It is generally known that CHCs have an important role in waterproofing mosquitoes [78], [86], [87]. Alterations in the CHC structure can also affect a mosquitoes resistance to desiccation. A 2010 study by Urbanski *et al* investigated the molecular physiology of *Ae. albopictus* eggs in terms of desiccation resistance [178]. Their study focused on the egg diapause initiated by photoperiods in temperate and tropical populations. This was investigated by profiling the HC composition on the surface of eggshells and quantifying a fatty acyl-CoA elongase which is known to be upregulated in shorter days [178]. They employed a GC-MS platform to quantify HCs on the eggshells although the data obtained was not shown. Urbanski *et al* reported higher surface HCs in temperate populations under shorter days, whereas tropical populations showed no difference [178]. None of the conditions showed a qualitative difference in the HC profiles [178]. Similar results were also reported for the fatty acyl-CoA elongase quantities [178]. This was attributed to the requirement for higher desiccation resistance in temperate populations due to the harsh winters. Interestingly, the temperate population showed a significant difference between shorter (higher HC) and longer (lower HC) days. Both days were lower than in tropical populations which showed no significant difference between shorter and longer days. Unfortunately, the results from temperate and tropical populations were not statistically compared to each other and so it is unknown if these differences are significant or not. Nevertheless, a clear link between HCs and desiccation is evident. As such, it was also studied by Arcaz *et al* in 2016, this time in *An. coluzzi*. Arcaz *et al* investigated the effects of thoracic spiracle size and CHCs on desiccation tolerance [179]. In their study, a mortality assay was used under dry and rainy season conditions. Each mosquito was weighed prior to the experiment. During the experiment, mosquitoes had no access to food or water. Dead mosquitoes were collected and weighed to assess water loss. Additionally, randomly selected dead mosquitoes were placed in a desiccation oven prior to dissection for spiracle size

measurement, followed by CHC analysis by GC-MS. There was no significant difference reported between the initial and end mass of mosquitoes [179]. Arcaz *et al* concluded that a higher total CHC abundance was not the sole determinant of desiccation tolerance, but that the abundance of specific CHCs played an important role in desiccation tolerance [179]. A caveat to this study is CHC profiling was performed on dead mosquitoes which were starved and not supplied with water. Under such conditions, the HC profile might have been altered due to extreme stress. For example, in desiccation conditions, mosquitoes can oxidise triglycerides to produce water [187].

Perhaps within the studies covered here, Hidalgo *et al* (2018) produced the most comprehensive study investigating differences between dry and rainy season condition in *Anopheles* species (*An. gambiae* and *An. coluzzi*). In order to probe the physiological and metabolic mechanism of the well-known aestivation process, Hidalgo *et al* used a combination of gas exchange measurements, gene expression analysis and HCs profiling. In total, three populations (two *An. coluzzi*, one *An. Gambiae)* were bred in dry and rainy season conditions [181]. The profiling of these mosquitoes was performed by measuring gas exchange (O_2 and CO_2), flight activity, allometric measures, gene expression and CHC fingerprints [181]. Due to the scope of this literature search, only the CHC profiling and its relation to the desiccation will be discussed. The CHC profile of each population was measure *via* GC-MS and a total of 13 compounds were identified and quantified [181]. Linear discriminant analysis showed a clear separation between the dry and rainy season populations. Although the loadings of the discriminant analysis were not discussed in detail, three metabolites can easily be seen from the loadings: 3-methylpentacoase, 13-methylhexacosane, and n-heptacoase [181]. When HCs and their properties are considered, alkanes are classed as the compounds with the highest waterproofing properties, although it is interesting to see methyl branched alkanes were influential in the discriminant analysis.

The remaining publications that were found during the literature search were more focused on the profiling of the CHCs in mosquitoes. An early study on *An. stephensi* in 1997 by Anyanwu *et al* used a GC approach to investigate the CHC profiles of four strains, where two were resistant and two were susceptible to DDT and malathion [84]. This study used GC chromatograms to profile CHCs and used a combination of PCA and linear discriminant analysis to model the differences between different experimental groups [84]. Surprisingly, the discriminants were not dissected further in order to determine the most influential HCs in the discrimination. The study successfully managed to model the resistance status by CHC

profiles although their genetic background could not be supressed [84]. It should be noted that, it was reported that all four strains were obtained from a laboratory, but how long these strains were maintained for was not reported. This might explain the interference of genetic background in the analysis. This study acknowledges the genetic background was a limitation and the fact that there are four population where half are resistant and the other half susceptible could have been investigated differently. Discriminant analysis could have been performed to accentuate differences between resistance status giving the study more confidence over the result. Furthermore, investigating the discriminants themselves would have given an insight into the key HCs in discriminating between different populations as well as resistance status.

A similar study was undertaken by Liebman *et al* in 2015, investigating the effects of diet on age determination in *Ae. aegypti* using near infrared spectroscopy (NIRS) [177]. NIRS is a lot less common compared to NMR and MS in metabolomics studies although typically Fourier-transform infrared spectroscopy (FTIR) is used. Liebman *et al* investigated the effects of diet in determining the age of *Ae. aegypti* mosquitoes. In order to do this, measurements from the head and thorax were taken without extraction *via* NIRS which measures the CH, NH and OH functional groups [177]. In their study, Liebman *et al* focused on correctly classifying mosquitoes with a cut off of 7 days old. This was selected since it takes 7 days for female *Ae. aegypti* to become infected with Dengue virus after an infected blood meal [177]. Due to the nature of NIRS which measures the absorbance of functional groups, it is not always possible, if at all, to identify individual CHCs, although the cumulative absorbance recorded from the CHC layer can be used to create a statistical model representing the CHC layer. Statistical models were built by either fitting a linear or polynomial model [177]. The highest performing model was a polynomial model with an R^2 of 0.6, which is not particularly high [177]. Liebman *et al* discussed that regardless of the poor fit of the model, the prediction results with a 7 days old cut off was satisfactory and should be further improved by acquiring more data [177]. Although this may be true, the data presented in this work suggests that the prediction model is highly variable. For each age group, the predictions vary by approximately ± 3-4 days. NIRS is a very portable method compared to NMR and MS and can have a great impact in field studies. However, the statistical approaches taken need to be revised to fully explore the suitability of NIRS in such applications although it is unlikely for NIRS to be a common application for CHC profiling on its own.

GC-MS is currently considered the current gold standard for CHC profiling. Two studies conducted in 2014 by Wagoner *et al* [180] and Reidenbach *et al* [182] both investigated changes in *Anopheles* species in dry and rainy seasons using GC-MS profiling. Wagoner *et al* focused on the water control mechanisms in these seasonal conditions whereas Reidenbach *et al* investigated the mechanisms through inversion of 2La chromosome. Wagoner *et al* hypothesised larvae maintained under dry season conditions (i.e. shorter days) would induce aestivation and would produce larger adults with smaller spiracle size in order to preserve more water [180]. In their research, Wagoner *et al* compared the wing size, spiracle size and CHC composition of larvae and adults under dry and rainy season conditions [180]. Dry and rainy seasons were simulated through different relative humidity as well as different length photoperiods. Wagoner *et al* reported wing size of adults in the dry season conditions were 4.4% larger [180]. Interestingly, these mosquitoes also took a shorter time to hatch compared to the shorter winged rainy season mosquitoes. When the spiracle sizes were compared, dry season mosquitoes showed a 5% increase in size which was unexpected as during dry season conditions, it would be expected for spiracle size to be smaller in order to have better control over water loss [180]. Mosquitoes lose water through respiration which happens through the spiracles. It can be speculated that larger mosquitos would require larger spiracles in order to have enough gas exchange for their body size, even though this may result in more water loss. Wagoner *et al* (2014) used a GC-MS approach to analyse CHCs and did not report any qualitative differences between the two seasons. On the other hand, when their quantities were considered, dry season mosquitoes had significantly more CHCs compared to rainy season mosquitoes [180]. A total of 15 CHCs were reported with only nine being identified (six alkanes, two alkenes and one dimethyl-alkane) [180]. Unfortunately, the CHC levels were not standardised to the body size and so it is unknown if this difference is due to seasonal conditions. A previous study by Urbanski *et al* (2010) [178] reported higher CHCs on *Aedes* eggs undergoing quiescens in drier seasons and so such CHC changes are likely to be seasonal. Another interesting observation reported was that regardless of the rearing condition the young virgin females did not present any difference in their CHCs. Thus, Wagoner *et al* concluded that CHC composition is predetermined during the larval and pupal stages and is not affected by the seasonal conditions [180]. In spite of pupal variation, CHC composition is dynamically altered in adulthood, especially between virgin and mated mosquitoes. This is most likely for mating purposes. A caveat to this study is that in order to simulate the dry and rainy seasons, both humidity and day lengths were altered. Although this creates a better representation of the seasons, it also introduces extra variations in the experiment with the changes of two variables rather than one. These variations should have

been taken into account either in the experimental design or acknowledged in the discussions as a limitation.

Reidenbach *et al* (2014) investigated the same seasonal effects on *An. gambiae* and *An. coluzzi* with 2La chromosomal inversion [182]. Chromosome inversion is the 180˚ rotation of a chromosome segment in its location and can occur naturally. By doing so, the information content in the original segment (2La) is unchanged, but the order of the information is inverted (2L+a) [182]. The segment 2La has been implicated to effect heat and desiccation tolerance in literature [182]. Reidenbach *et al* measured the water lost between the karyotypes 2La and 2L+a under two seasonal conditions and reported significantly higher water loss in 2La karyotype mosquitoes [182]. Further extrapolating on the desiccation resistance properties, the CHC profiles were analysed in order to reveal if any link between 2La karyotype and CHCs is present. Reidenbach *et al* reported sixteen CHCs (nine alkane, two alkene, four methyl-branched alkanes, and an unidentified) with no qualitative difference between karyotypes or seasonal conditions [182]. In arid conditions, unbranched alkanes were preferred over branched alkanes in either karyotype. Additionally, in the inversion karyotype (2L+a), a greater abundance of unbranched saturated alkanes was observed [182]. The 2L+a karyotype showed a higher abundance of alkanes compared to the 2La karyotype, irrespective of seasonal conditions [182]. Reidenbach concluded that the inverted 2L+a karyotype gives higher waterproofing properties to the mosquito. Reidenbach also noted that this inversion is rare in desiccation resistant species suggesting water control mechanisms might be more complex and have multiple compensation mechanisms associated with them.

The last study on CHC profiling found in the literature search was the 2016 study by Hamid *et al* investigating the behavioural response to CHCs in *Ae. aegypti*. Using GC-MS, Hamid *et al* identified n-heptacoase as the major constituent of the CHC extract *via* hexane [176]. Within this extract, all reported compounds were long chain alkanes. In this behavioural study, mosquitoes' attraction to CHCs of different ages were tested [176]. The experimental design tested male and female hexane extract on the same sex and the opposite sex. Hamid *et al* reported that both males and females were more attracted to the older members of their own sex, while for the opposite sex, both were most attracted to three day old members [176]. An interesting result of this study is that previously alkanes were typically not referred to as compounds with communication properties [Chung and caroll] and so it suggests that perhaps long-chain alkanes could be used in communication as well as waterproofing in

mosquitoes. If so, given their more stable nature compared to pheromones, n-alkanes might be used in the communication of more static information such as age, sex, and maturity rather than information of a more dynamic nature such as surroundings and potential danger. Another explanation might be that the minor constituents of the hexane extract contain semiochemicals used in communication which are very potent or more simply that long chain alkanes have more than one property which has not been fully characterised yet.

From the studies covered in this section, there is a clear link between CHCs and their role in resistance *via* reducing the insecticide penetration. Furthermore, there is strong evidence in the Cyp4g family, especially Cyp4g16, for their role in hydrocarbon production as decarbonylases. Several studies showed strong evidence of a higher abundance of total CHCs in desiccation and/or insecticide resistant species, particularly the alkanes and methyl branched alkanes. Interestingly, no study investigated the precursor of these mechanism, thus creating a knowledge gap.

Considering the studies covered in section 1.6.6 and here, three topics stood out as areas for further investigation.

1) The lack of metabolomic studies between males and females in mosquitoes should be addressed. Establishing the metabolic differences between sexes can yield important information on how experiments should be designed (e.g. pooling of males and females).
2) A metabolomics polar approach into increasing the understanding of the relationship between CHCs and the Cyp4g16 and Cyp4g17 enzymes.
3) A comparison study of the polar metabolites of resistant and susceptible species.

The work presented in this thesis will try to fill these knowledge gaps.

1.7 Aims and objectives

The overall objective of this work was to use NMR metabolomics to bring further knowledge on insecticide resistance and differences in metabolism between the sexes in mosquitoes. More specifically, Cyp4g16, and Cyp4g17-related mechanisms thought to be involved in cuticular resistance through pupal and adult life stages were investigated. Furthermore, wild type resistant and susceptible mosquito strains were compared in order to deepen the understanding on resistance in wild type mosquitoes. This project is based on the hypotheses that (i) the different life stages of mosquitoes have distinct metabolic needs, and (ii) the development of resistance through enzymatic alterations will manifest changes in metabolic pathways that can be characterised through NMR metabolomics. This entirely novel research investigation is carried out in close collaboration with Dr. Gareth Lycett at the Liverpool School of Tropical Medicine.

Anopheles and *Aedes* species are vectors of 31% of the closely monitored diseases by WHO. *Anopheles* species account for the highest mortality, and the *Aedes* species account for the greatest number of diseases contracted. Within these genera, *An. gambiae* and *Ae. aegypti* have been identified as the most important vectors and as such were selected for this project. Furthermore, in order to explore CHC involvement in resistance, the effects of Cyp4g16 and Cyp4g17 KDs were selected as paralogues of Cyp4g1 found in *D. melanogaster*. In this project, NMR was selected as the primary analytical technique as mosquitoes typically weigh around 1-2 mg providing enough material for measurment, with NMR providing a very robust and reproducible pipeline where the majority of the data acquisition can be automated which aids with consistency.

The outline of the following chapters are:

In chapter 2, the methods employed in the metabolomics framework that was developed are explained in detail, covering mosquito rearing to data analysis.

In chapter 3, metabolic differences between sexes in *An. gambiae* and *Ae. aegypti* species across pupal and adult stages will be determined. In order to accomplish this, a metabolomics framework will be built based on previous methods published on metabolite extraction, data acquisition and data analysis. This is a novel study and as such a considerable amount of time

was spent on optimising the sample preparation, data acquisition and specifically data analysis protocols.

In chapter 4, mosquitoes carrying individual gene knock-downs of Cyp4g16 and Cyp4g17 from pupal and adult stages will be compared using the metabolomics framework established. Both Cyp4g16 and Cyp4g17 are thought to catalyse the same decarbonylation step, although both enzymes have distinct secondary structures and exhibit different temporal expression profiles at early larval stages. This chapter aims to increase the understanding of the differential function of these genes through analysis of metabolite disruption in the mutant mosquitoes and the relevant metabolic pathways affected.

In chapter 5, wild type mosquito strains resistant to insecticides are compared to susceptible strains in both sexes using the metabolomics framework. The strains tested are *An. gambiae* VK7 (resistant) and N'Gusso (susceptible) and *Ae. aegypti* Cayman (Grand Cayman; resistant) and New Orleans (USA; susceptible). A caveat to these studies is that these strains do not have close intraspecies genetic backgrounds which may interfere with creating a more complex analysis. The analysis is performed at pupal and adult stage to probe the resistance profile of wild type mosquitoes.

In chapter 6, a discussion of the results obtained in chapters 3-5 will be carried out in relation to the literature and overall conclusions are drawn. Additionally, further improvements, future work and applications of the findings are discussed. Finally, this work has been critiqued in terms of assumptions, limitation, and disadvantages.

Chapter 2

2 Materials and methods

2.1 Mosquito rearing

All mosquitoes were reared in the Liverpool School of Tropical Medicine (LSTM) under the supervision of Dr Gareth Lycett. For all rearing the following conditions were used: temperature: 27 ˚C, humidity: 75%, 12:12 hour:hour night and day cycle with dusk and dawn simulations (30 min each). All mosquitoes were reared from established colonies (100% clonal adaptation).

2.1.1 Floating eggs

Eggs were laid by female mosquitoes on a filter paper (Whatman, UK) after a blood meal (*An. gambiae*: fresh blood donated by Dr Gareth Lycett, *Ae. aegypti*: 50% red blood cells in blood plasma (National blood service, Liverpool) with Haemotek membrane feeding system: SP4W2 (Haemotek; Blackburn, UK). These eggs were floated on trays with tonic salt water (TSW). To prepare TSW, Pond Guardian Tonic Salt (Interpet; Surrey, UK) was dissolved in reverse osmosis (RO) water (1 gL^{-1}). A plastic tray (7.5 x 25 x 25 cm) was filled with TSW up to 1 cm in height. Filter paper containing the eggs was then washed with TSW using a wash bottle over the tray and the filter paper left floating on the tray. To prevent contamination, the tray was covered with a net. Eggs and the filter papers were washed twice a day with TSW using a wash bottle until the eggs hatched.

2.1.2 Larval stage

When eggs hatched, the filter paper was removed from the trays. The trays were washed daily with TSW and all the larvae were fed. *An. gambiae* larvae were fed with TetraMin fish food (Tetra; Melle, Germany) and *Ae. aegypti* larvae were fed with 300 mg of brewers yeast (Natures Aid; Huddersfield, UK). Trays containing more than 200 larvae were split into two trays. This process was repeated until all larvae hatch into pupae.

2.1.3 Pupal stage

All pupae samples were collected within 12 hours of pupating. Pupae samples were collected and placed in bowls containing fresh TSW. The collected samples were sex classified and screened (if required) under a microscope (Figure 2.1-1). Each pupa was then transferred

into a new Eppendorf with 500 µL of TSW to wash off the dirty water from the larvae tray. After this washing step, the TSW was removed and samples were flash frozen in liquid nitrogen. Collected samples were stored at -80˚C until extraction.

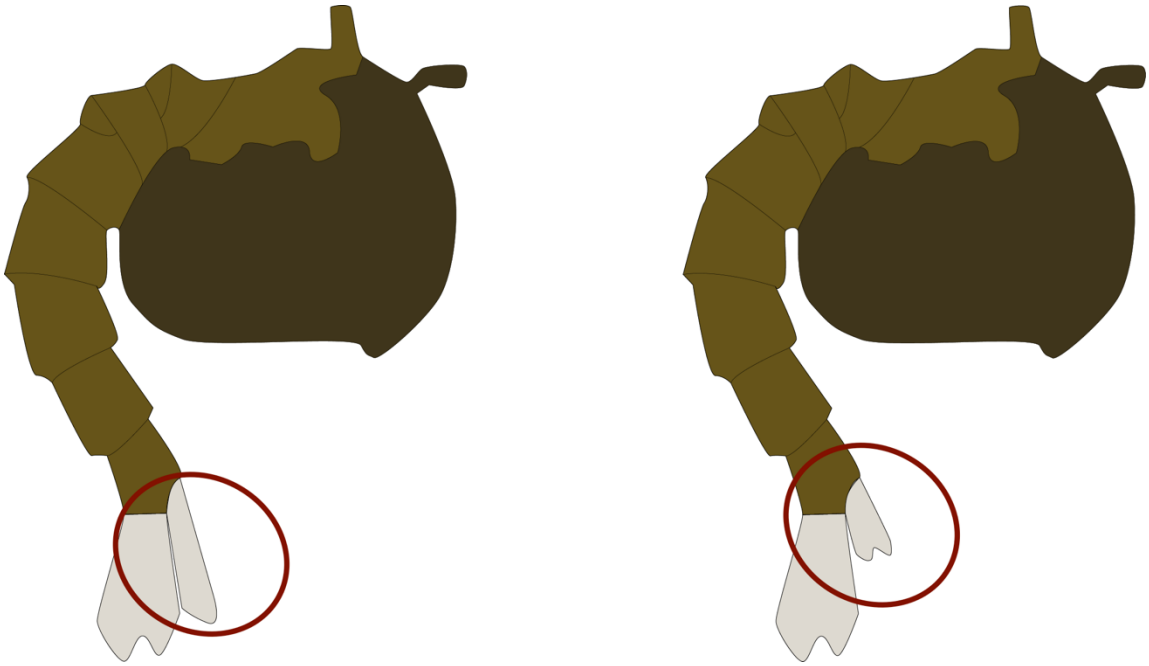


Figure 2.1-1: Comparison of male (left) and female (right) pupa illustrations. Sexing pupa can easily be achieved by comparing genitals (red circles) by the anterior fins.

To rear adults, pupae were collected within 12 hours of hatching and placed in bowls with fresh TSW. These bowls are then placed in buckets and covered with netting. Damp cotton wool soaked in 10% sucrose (Tate and Lyle, UK) dissolved in TSW was placed in the bucket; this provided the feed for the maturing mosquitoes.

2.1.4 Adult stage

Mosquitoes that emerged were collected *via* standard mouth aspirator within 12 hours of emergence. After collection with the aspirator, mosquitoes were anesthetised using CO_2 (5 $Lmin^{-1}$ flow rate). Anesthetised mosquitoes were sexed visually and placed in individual Eppendorf tubes *via* fine point forceps. Upon waking (typically 5 minuntes after aesthesia), mosquitoes were checked prior to collection. Only mosquitoes that were intact and capable of flying were collected for analysis. Collected samples were flash frozen in liquid nitrogen and stored at -80 ˚C until extraction.

2.1.5 An. gambiae knock-down crossing and Gal4/UAS screening

The generation of the Gal4/UAS (Figure 2.1-2-A) transgenic mosquito lines was performed by Dr Gareth Lycett and a detailed explanation of the procedure can be found in Lynd *et al* 2019 [188]. Briefly, an oenocyte specific Gal4 driver line was created by insertion of a promoterless Gal4 gene into the *An. gambiae* genome close to a suspected oenocyte enhancer. The Gal4 construct was marked with a red fluorescent protein gene. Suitable UAS vectors to induce RNAi of each gene were also constructed, in this case using an eYFP (yellow fluorescent protein) marker gene. In order to create the Cyp4g16 and Cyp4g17 knock-down constructs, fragments of the respective *An. gambiae* genes from a G3 strain were cloned. These fragments were placed as two copies in opposite orientations, separated by their native introns to create templates that would be transcribed and spliced to produce hairpin RNA interference (RNAi) constructs. These hairpin fragments were cloned downstream of the UAS domain to create the final RNAi vectors.

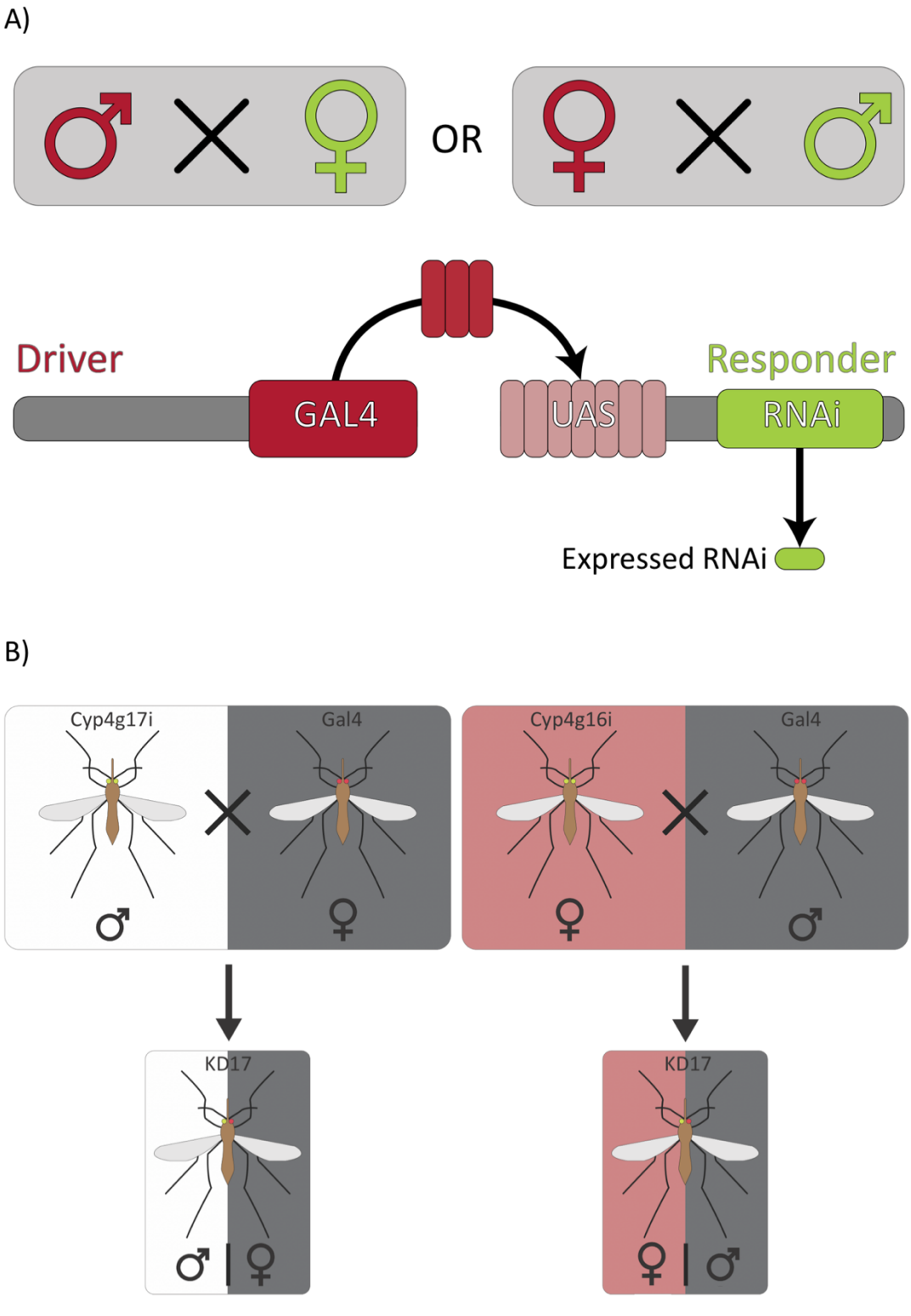


Figure 2.1-2: A) Gal4/UAS system requires a line with a driver and a line with a responder. Upon breeding, some progenies will have both the driver and responder lines. When they are present together, the driver line promotes production of Gal4 (red fluorescens) which is recognized by the UAS system (yellow red fluorescens). The activated UAS system will promote the expression of the gene on the responder line. This can be a gene to overexpress a protein or knock-down a gene, such as RNA interference (RNAi) for knock-downs. B) Crossing of Gal4 driver and UAS responder lines to produce mosquito eggs with knock-down mosquitoes.

A PhiC31 vector was used to transform two different docking strains of *An. gambiae* embryos to generate the driver and responder lines. For Gal4 driver lines, early embryos from transgenic attP docking line A14, in which the docking site was under the transcriptional influence of an oenocyte enhancer, were co-injected with the Gal4 donor plasmid and PhiC31 helper plasmid [188]. Similarly, for generation of cyp4g16-RNAi and cyp4g17-RNAi, the respective UAS plasmids were co-injected with helper plasmid into embryos from A11 line [188]. Following the injections, mosquitoes were reared under standard conditions and mated to wild type mosquitoes, as explained in section 2.1. Larval progeny from these matings were screened for the fluorescent eye markers to identify individuals carrying stably inherited driver (red fluoresence) and responder (yellow fluoresence) constructs. These were interbred to create populations of each of the transgenic driver and responder lines.

In order to activate the knock-downs of Cyp4g16 and Cyp4g17, pupae were collected from the established transgenic lines for sexing, screening and crossing. Sexing and screening for the Gal4 driver (RFP) and UAS responder (YFP) was performed using a Leica MZFLIII microscope. Figure 2.1 2 shows the crossing performed in order to knock-down Cyp4g16 and Cyp4g17. Upon sexing and screening, up to 50 pupae from each of following sexed strains were selected: male Gal4, female Gal4, male Cyp4g17 responder and female Cyp4g16 responder, and crossed to produce progeny that expressed the respective RNAi hairpins in an oenocytes specific manner (Figure 2.1-2-B).

2.2 Metabolite extraction

All extractions were performed on samples retrieved from -80 ˚C storage. Due to the residual water in pupae samples, thawing of the residual ice was performed over ice ensuring a uniform interaction of the solvent with the sample. To keep the extraction protocol consistent adult samples were thawed over ice prior to extraction in the same manner. All solvents used in the extraction procedure were ice-cold. Except for the different solvent composition, both polar and non-polar extraction procedure follows that same protocol shown in Figure 2.2-1.

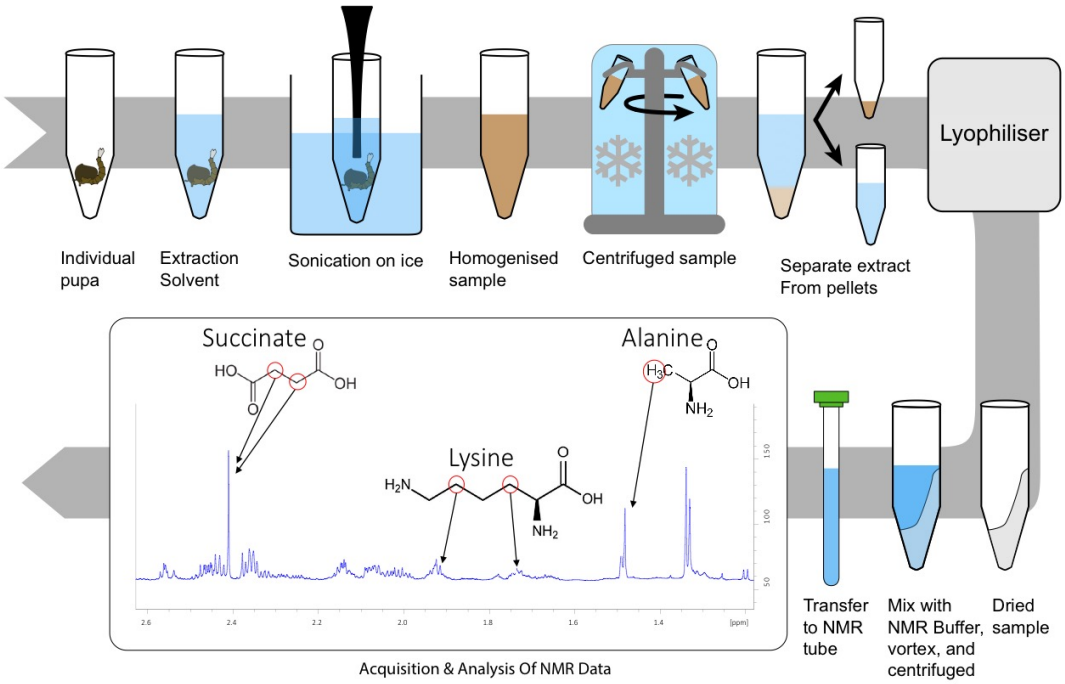


Figure 2.2-1: A typical extraction method starting from a fresh/stored sample. Addition of the extraction required solvent (polar or non-polar) followed by an incubation over ice ensured solvent penetration through the sample. Homogenisation was critical for the separation of protein and small molecules. Homogenised samples were centrifuged to separate the debris and precipitants (this allowed the solution of metabolites to be extracted). The supernatant was lyophilized and metabolites either stored or prepared for analysis. Dry pellets were mixed with the appropriate buffer depending on the technique (Mass spectrometry or NMR) to be used. In this illustration the technique was NMR.

Extraction method of polar metabolites was adapted from Beckonert *et al* [189]. For polar metabolite extraction, a 50% acetonitrile (HPLC grade Fisher scientific) solution in RO water was used. Each sample was mixed with 500 µL of 50% acetonitrile and incubated on ice for 10 minutes. Samples were then homogenised using a sonicator (MSE Soniprep 150 plus) fitted with an exponential probe operating at a frequency of 23 KHz and amplitude of 10 µm. Each sample was sonicated with 3 x 30 s pulses and 30 s breaks in between pulses over an ice bath. Homogenised samples were then centrifuged at 21,000*g* for 10 minutes at 4 ˚C to separate the non-soluble material. Aliquots from the centrifuged samples were transferred to a fresh microtube. Each was flash frozen in liquid nitrogen prior to lyophilisation. Frozen samples were lyophilised (Thermo Scientific, Power dry CL 3000) at -56 ˚C for 16 h. Lyophilised samples were stored at -80 ˚C until NMR sample preparation.

Sample retrieved from -80 ˚C storage were mixed with an NMR buffer solution comprised of 100 mM sodium phosphate buffer (Sigma-Aldrich), 0.1 mM 3-(Trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt (TSP) (Sigma-Aldrich), 1.2 mM sodium azide (Sigma-Aldrich) in ^2H_2O (Sigma-Aldrich) was used. Lyophilised samples were mixed with 200 µL of the NMR Buffer. Each sample was then vortexed and centrifuged at 12,100*g* for 2 minutes. Samples

were then transferred to NMR tubes (diameter: 3 mm, length: 103.5 mm & Borosilicate glass).

2.3 NMR setup and data acquisition

All samples were acquired using a Bruker Ascend 700 MHz spectrometer fitted with a 5 mm TCI Cryoprobe, Avance III HD console and a SampleJet automated sample changer with a chiller set to 4 ˚C.

2.3.1 Temperature calibration

NMR data is sensitive to temperature fluctuations. Although the temperature can be kept stable in the magnet, environmental temperature can cause an offset in the temperature setting process. To minimise environmental temperature fluctuations, NMR spectrometers are kept in a temperature controlled environment set to 21 ˚C.

Changes in temperature causes systematic shift in an NMR spectrum. These shifts are on a linear scale by nature hence can be exploited to measure the actual sample temperature once it is inserted into the probe [190]. By measuring the distance between two peaks in a 99.8% Methanol-d_4 sample (Bruker, product code: Z10627) the actual sample temperature sitting on the probe was calibrated with an accuracy of ±0.1 ˚C [190].

2.3.2 Shimming

NMR data acquisition requires a homogenous magnetic field in the three-dimensional space. When a sample is inserted into the magnet, it interferes with the homogeneity of the magnetic field. Between samples of similar nature, the magnetic field on the Z-axis distorts the most, whereas X and Y axes fields are more stable. Prior to each dataset, magnetic field homogeneity was calibrated using a standard sample comprised of 2 mM Sucrose, 0.5 mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS), 2 mM sodium azide in 90% $^{1}H_2O$ and 10% $^{2}H_2O$ with 40 mm filling height (Bruker, product code: Z10902). Following calibration, the TSP signal was used to assess the homogeneity of the magnetic field. Overall peak shape and half height full width (HHFW) was checked (less than 1 Hz).

2.3.3 Water suppression

Water is the most common solvent for NMR in biological samples. Due to the high abundance of ^{1}H isotope using water gives a very intense signal that saturates the receiver hence,

masking signals coming from metabolites with low concentrations. In order to overcome this problem all samples were dissolved in NMR buffer prepared in 99.9% $^{2}H_2O$. Even though ^{2}H is not detected by NMR, exchange with ^{1}H and trace amounts of $^{1}H_2O$ can still cause signal overflow and/or distortions around water region (approximately 4.7 ppm). This was resolved by using NMR solvent suppression methods. Two most commonly used were used presaturation and excitation sculpting were used for 1D and 2D experiments respectively.

2.3.3.1 Presaturation

Presaturation suppression is applied by irradiating the solvent resonance with long low-powered selective pulse before the experimental pulse sequence is initiated. By applying this selective pulse, the water spins are flipped on to the X-Y plane while the spins arising from metabolites are still on Z-axis. By applying the first hard 90˚ pulse of the experiment, all metabolite spins are flipped on to the X-Y plane where acquisition can take place. This first hard 90˚ pulse also flips the water spins which were on the X-Y plane to Z-axis where no acquisition occurs. Therefore, supressing the signals originated by the ^{1}H in water. In this project the presaturation was applied with a power level of 49.98 dB for 20 µs.

2.3.4 NMR data acquisition

Throughout this project 4 NMR experiments were used to generate data. 1D-Nuclear Overhauser spectroscopy (NOESY) were used for checking the shims and water suppression per sample during data acquisition. Carr-Purcell-Meiboom-Gill (CPMG) experiment was used to measure metabolites for statistical analysis. ^{1}H-^{1}H total correlation spectroscopy (^{1}H-^{1}H TOCSY) and ^{13}C-^{1}H-heteronuclear single quantum coherence (^{13}C-^{1}H HSQC) for complementary information on metabolite identification.

2.3.4.1 1D-Nuclear Overhauser spectroscopy (1D-NOESY)

1D-NOESY experiment is a ^{1}H experiment where both large and small molecules can be observed. It is based on a 2D experiment where coupling between ^{1}H-^{1}H is observed through space. The 1D implementation of this experiments uses presaturation and gradient pulses to achieve water suppression. In this project this experiment was utilised as a quality check during data acquisition as well as a critical part of the automation process for high-throughput. The automation program integrated to the 1D-NOESY experiment was used to save common experimental parameters to be transferred to the subsequent experiments such as: CPMG, ^{1}H-^{1}H TOCSY and ^{13}C-^{1}H HSQC.

2.3.4.2 Carr-Purcell-Meiboom-Gill (CPMG)

CPMG is a 1D ^{1}H experiment where larger molecules are filtered out by exploiting the relaxation properties of molecules. When excited with a pulse, over time molecules will 'relax' (T1 relaxation) and recover their pre-excited state on the Z-axis (along the magnetic field). Relaxation time and molecule has an inverse relationship, where larger molecules relax faster, and smaller molecules relax slower. In a CPMG experiment using carefully timed pulses, molecules are 'locked' on the X-Y plane (also known as spin echo lock) for a predefined time. While 'locked' relaxation takes place and larger molecules relax quicker than smaller ones. When the signal acquisition starts larger molecules are effectively filtered out while smaller molecules remain.

In this project CPMG experiment was used as the main source of data for statistical tests. It was preferred over 1D-NOSEY due to the residual large molecules that could not be separated *via* physical methods (i.e. extraction and centrifugation).

2.3.4.3 ^{1}H-^{1}H Total correlation spectroscopy (^{1}H-^{1}H TOCSY)

TOCSY is a homonuclear experiment where correlations within an unbroken chain of coupled spins are observed. TOCSY is an extension of the experiment correlation spectroscopy (COSY) where the coupling between ^{1}H-^{1}H through 3 bonds is observed. Unlike COSY, TOCSY can observe couplings four of more covalent bonds away. For example, if ^{1}H-A is coupled to ^{1}H-B and ^{1}H-B to ^{1}H-C while ^{1}H-A is not directly coupled to ^{1}H-C, TOCSY would show a correlation both from ^{1}H-A to ^{1}H-B and ^{1}H-A to ^{1}H-C. In TOCSY experiments, the covalent bond coverage depends on a parameter called 'mixing time' where longer time is required for higher number of bond correlations. The trade of is that higher mixing time results in longer acquisition time and a reduction in sensitivity due to relaxation of the molecules. ^{1}H-^{1}H TOCSY experiments were used to resolve ambiguity in 1D-NMR identification due to overlapping. For all TOCSY experiments a mixing time of 80 ms were used.

2.3.4.4 ^{13}C-^{1}H Heteronuclear single quantum coherence (^{13}C-^{1}H HSQC)

HSQC is a heteronuclear correlation experiment typically in between ^{13}C-^{1}H or ^{15}N-^{1}H. HSQC achieves heteronuclear correlation through transferring of polarisation from highly abundant NMR active nuclei (typically ^{1}H) to lower abundant NMR active nuclei (e.g. ^{13}C). This is called insensitive nuclei enhanced by polarisation transfer (INEPT). Following the transfer of

polarisation from 1H to 13C, the sensitivity enhance polarisation is transferred back to 1H for direct measurement. By doing so chemical shifts of 13C are indirectly recorded on the directly measured 1H dimension. An 13C-1H HSQC shows chemical shifts of protons and carbon that are directly bonded. Benefiting from the large chemical shift scale of ^{13}C (0-200 ppm), overlaps observed in ^{1}H spectra (due to low chemical shift range of 0-12 ppm) can be resolved further.

2.3.4.5 NMR experiment parameters

NMR experiments were acquired as following: ^{1}H 1D-NOESY (48 K complex points, 32 scans, 4 dummy scans) and 1D with ^{1}H 1D-CPMG pulse (48 K complex points, 128 scans, 4 dummy scans) experiments were acquired for each sample. Further details of the experiments are shown in Table 2.3-1.

Table 2.3-1: Acquisition parameters for NMR experiments.

Experiment	1D-NOSEY	CPMG	^{1}H-^{1}H TOCSY	^{13}C-^{1}H HSQC
Pulse program	noesygppr1d	cpmgpr1d	dipsi2esgpph	hsqcetgpsisp.2
Complex points[(a)]	^{1}H = 48 K	^{1}H = 48 K	^{1}H = 2048 ^{1}H = 512	^{1}H = 1024 ^{13}C = 512
Number of scans	128	128	16	32
Dummy scans	4	4	32	16
Temperature [˚C]	25	25	25	25
Acquisition time [s]	2.7262	2.7262	^{1}H = 0.1218 ^{1}H = 0.0304	^{1}H = 0.0731 ^{13}C = 0.0121
Spectral width [ppm]	25.75	25.75	^{1}H = 12.0018 ^{1}H = 12.0018	^{1}H = 10.0015 ^{13}C = 119.8242
Receiver gain [rel]	79.66	79.66	169.84	169.84
Dwell time [µs]	27.73	27.73	59.5	71.4
Relaxation delay [s]	4	4	2	1.5
Mixing time [ms]	10	NA	80	NA

[(a)]K is equal to 1024.

2.3.5 NMR spectra processing

Acquired spectra were processed using standard automation routines for consistency. Spectra were first zero filled 128 K data point (K = 1024), followed by standard Bruker phase correction method apk0.noe where an exponential window function with 0.3 Hz line broadening is applied prior to Fourier transformation. Lastly, spectral referencing set as such that TSP signal is at 0 ppm. No baseline correction was applied during processing.

2.3.6 NMR spectra quality control and binning

Spectra were individually checked for quality control (QC). During QC, spectra were checked for overall flat baseline, width of the water residue (less than 0.25 ppm), half height full width of TSP peak (less than 1 Hz), presence of TSP ^{29}Si satellite peaks and overall shape of TSP (Figure 2.3-1). Spectra passing the QC were used for metabolite identification and statistical analysis.

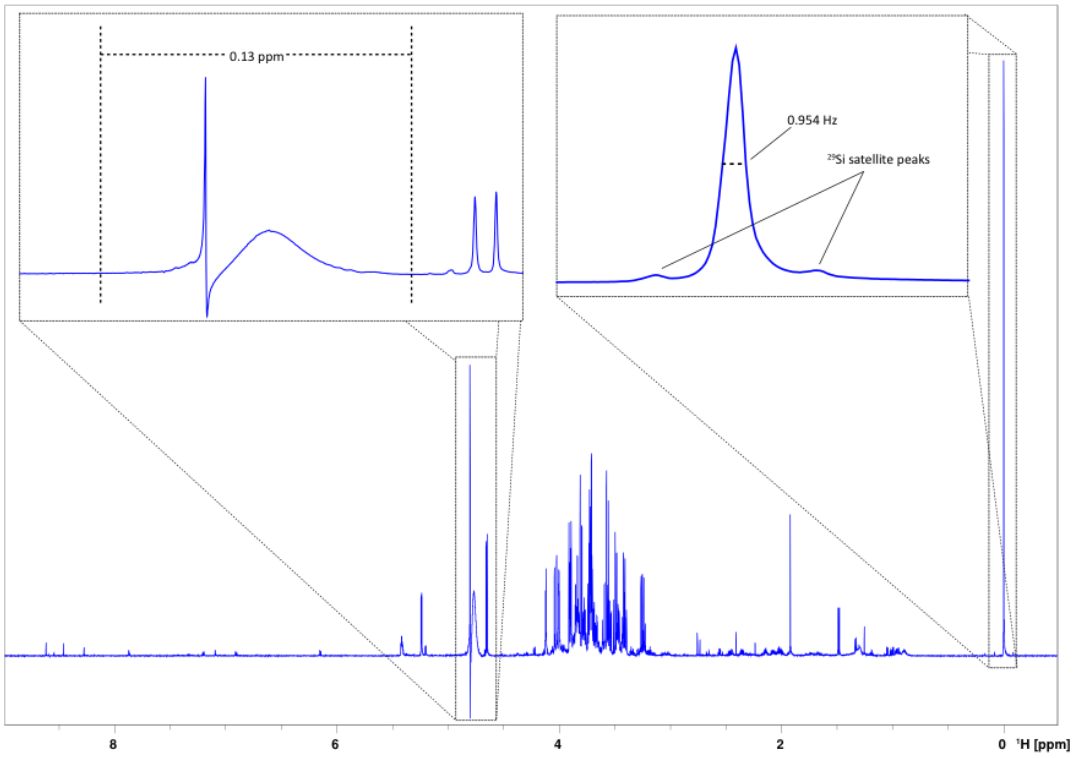


Figure 2.3-1: Proton NMR spectrum from a knock-down *An. gambiae* pupa satisfying the QC criteria. The spectrum demonstrates an overall flat baseline, symmetrical TSP (0 ppm) signal with a HHFW less than 1 Hz with satellite peaks arising from ^{29}Si in TSP and finally a water residue (4.5-5 ppm) less than 0.25 ppm.

In an NMR spectrum, the area under the peak is directly proportional to the concentration of the compound. In order to convert NMR spectra into a dataset that can be analysed with statistical methods, the peaks in the spectra are integrated. This process is called spectral binning and can be performed either by a set interval (incremental binning) or by manual selection (variable sized binning). In this project, manual selection was used in spectral binning. Each bin was defined by their chemical shift boundaries (in ppm), non-overlapping multiplet signals were binned together. Spectra were integrated using the defined boundaries *via* AMIX (Bruker, Coventry). Following peak integration, each peak was divided

by the peak region. During the binning process the water region (around 4.7 ppm) was excluded.

2.4 Metabolite assignment

Metabolite assignment is one of the critical processes in metabolomics as it is essential to know metabolite identities in order to find biomarkers and perform pathway analyses. The current acknowledged metabolite assignment standard is set by the metabolomics standards initiative (MSI) [191]. Although this method works well with MS techniques, it does not capture some of the intrinsic features of NMR. Table 2.4-1 shows an adaptation of the MSI scoring system in order to reflect NMR features.

Table 2.4-1: MSI assignment confidence levels adapted for NMR metabolomics.

Confidence level	Description
Level 1	Identified metabolites. This level requires two or more orthogonal properties of a standard compound to be analysed in the same laboratory where the experiment was conducted.
Level 2a*	Identified metabolites. This level requires matching one property of a standard compound analysed in the same laboratory where the experiment was conducted.
Level 2b*	Putatively identified metabolites. In contrast to identification, level 2a does not require the standard information to originate from the same laboratory.
Level 3	Putatively characterised compound classes. Same as level 2 annotation, it is used when the molecule can only be annotated as a particular class rather than a specific metabolite.
Level 4	Unknowns. Peaks that can be reproducibly detected and quantified.

*: Only for metabolites with NMR signature comprised of multiple signals.

Acquired spectra were assigned to individual metabolites using of the Chenomx software (Chenomx, CA) and an in-house library of metabolite spectra. Chenomx-identified metabolites were checked with in-house 1D standards (where available) to improve assignment confidence. Metabolite identification was also carried out by using 2D in-house spectra (^{1}H-^{1}H TOCSY and ^{1}H-^{13}C HSQC). Confidence scoring was performed according to the MSI scoring system where level 2a and level 2b are assignments done by internal and external databases respectively. Level 1 is standards acquired using the same equipment and further confirmed by an orthologous method (^{1}H and ^{13}C nuclei of NMR spectrum).

2.4.1 Level 1 assignment

Use of only ^{1}H information cannot yield a level 1 identification. For a metabolite to be identified with level 1 confidence, ^{1}H information needs to be complemented by a secondary orthologous method that annotates the same compound. In the context of NMR, use of ^{13}C

spectral information supplies the secondary orthologous information. NMR experiments such as ^{1}H-^{13}C HSQC (Figure 2.4-1), provides assignment information for both ^{1}H and ^{13}C nuclei hence yielding an identification. Unfortunately, due to the technical limitation of ^{13}C NMR, this information may not be possible to acquire for all concentration ranges.

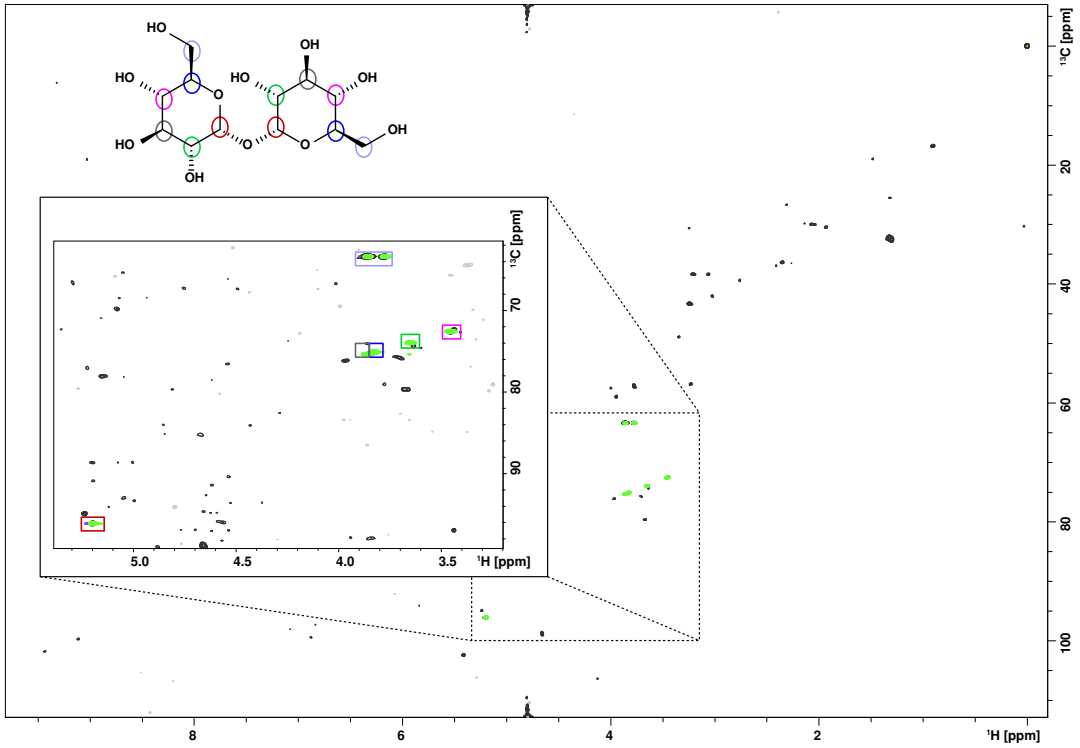


Figure 2.4-1: Identification of trehalose via ^{1}H-^{13}C HSQC. Using HSQC, both peaks arising from the hydrogen and carbon chemical groups the metabolite can be assigned hence providing further confidence in the assignment.

2.4.2 Level 2 assignment

Level 2a and 2b assignment is done by matching 1D-^{1}H NMR spectrum of a standard to the experimental spectrum (Figure 2.4-2). This process is done in two steps. The first step is the use of in-house metabolite libraries. These libraries are generated by acquiring the NMR spectrum of the metabolites on the same instrument where the experimental data was acquired. This type of annotation yields a MSI level 2a annotation. The second step is performed by Chenomx software (Chenomx, CA) which is designed to match 1D-^{1}H NMR spectrum of metabolite standards to an experimental spectrum using various matching algorithms. Assignment done *via* Chenomx yields a level 2b assignment, due to the use of external libraries.

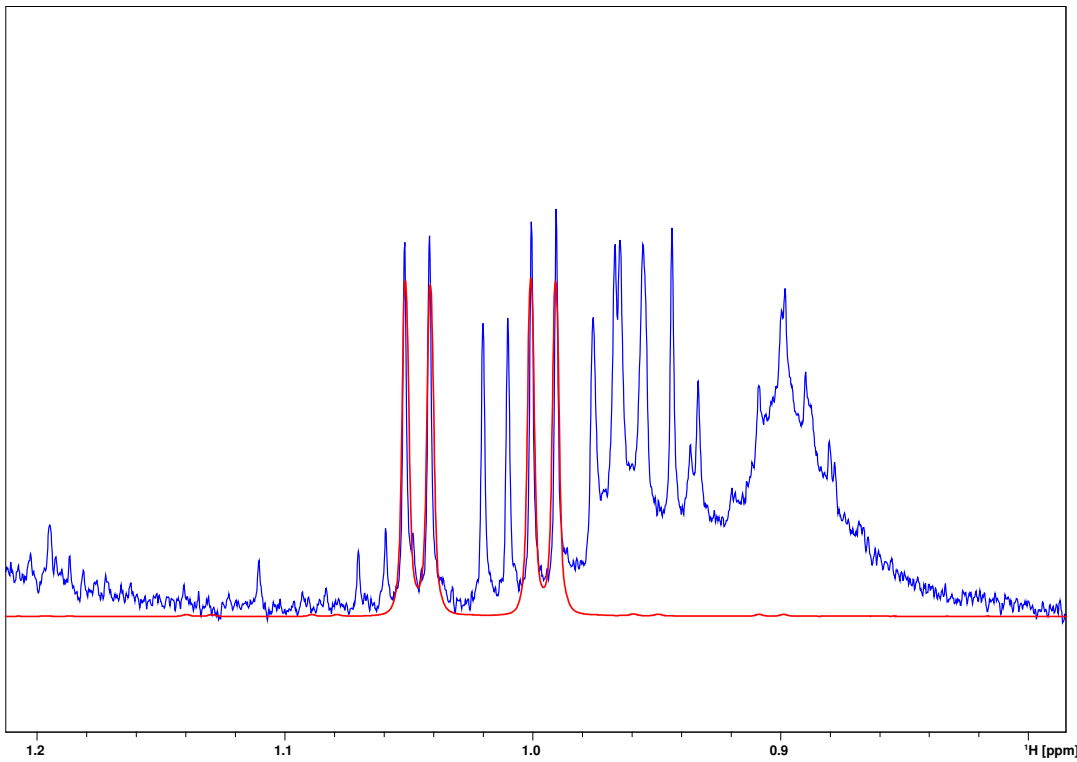


Figure 2.4-2:Matching of valine peaks from an in-house standard (red spectrum) with peaks on the mosquito extract experimental data (blue spectrum). Sample: Female KD16 pupa using in-house compounds yielding a level 2a annotation.

Compounds for the in-house libraries comprise of a mix of compounds classed as amino acids, sugars, purine and pyrimidines and carboxylic acids. Chenomx comes with a pre-built metabolite library for a range of magnetic fields (including 700 MHz). The Chenomx compound matching algorithm was designed to work with sample pH ranging from 4 to 9 and experiments acquired at 25 ˚C.

2.5 Statistical analysis

A flowchart of the statistical analysis is shown in Figure 2.5-1. Prior to the analysis, each NMR spectrum was subjected to quality control (QC) as explained in section 2.3.5. Any sample failing QC was feedback into the acquisition pipeline to acquire a new spectrum to a maximum of three times per sample, when failed samples were discarded from further analyses. This was done with the aim of retaining as many high-quality samples as possible. After QC, spectra were collated into a dataset for analysis. All further analytical steps were undertaken using a combination of published and custom-made scripts in R [192], [193] (Table 2.5-1). Statistical analysis was performed using R statistical package [192], [193]. Batch variation was assessed using principle vector component analysis (PVCA) [145]. The dataset was then normalised using probabilistic quotient normalisation (see section 2.5.1.2). Batch

effect correction was then applied as appropriate and data was scaled by standard deviation (autoscaling). The scaled dataset was subjected to multivariate analysis using a combination of principle component analysis (PCA) and partial least square – discriminant analysis (PLS-DA) (see section 2.5.2). Variable importance of the projection (VIP) was used to select important features from the PLS-DA models. Representative bins for the selected features were identified *via* correlation reliability score, (CRS, see further description in section 2.5.2.3.2). Verification PCA and PLS-DA models were built with the selected metabolites in order to demonstrate the information retained with the selected metabolites was informative of the groups to discriminate. Selected metabolites and differences between groups were used in suitable univariate tests (t-test/ANOVA-Tukey's HSD depending on number of groups to contrast). Finally, pathway analysis was performed *via* metabolite set enrichment analysis (MSEA).

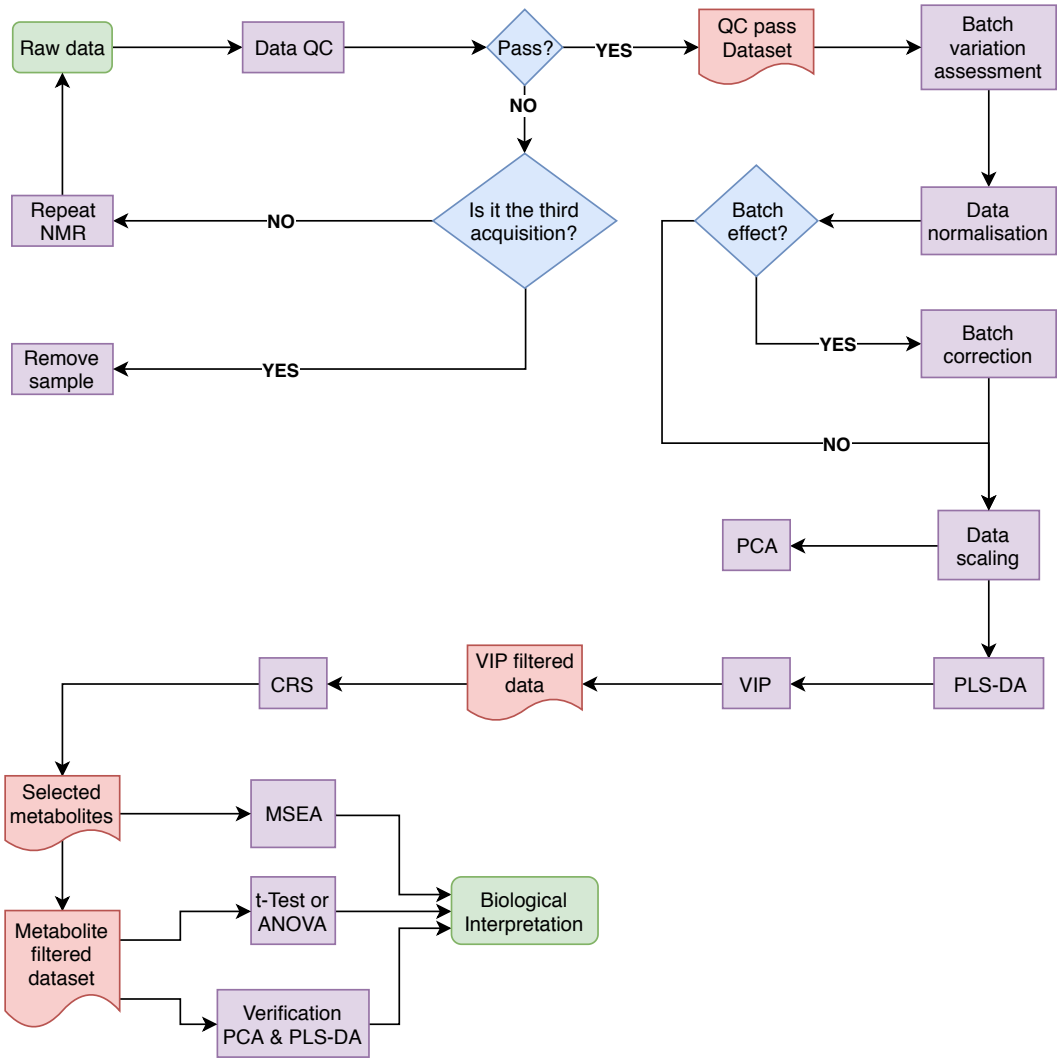


Figure 2.5-1: Metabolomics statistical analysis flowchart. Green rounded rectangles show start and end points. Purple rectangles show processes. Blue diamonds represent decision points. Red documents represent curated dataset.

All the data processing and statistical analysis was performed using R and Python statistical package [192], [193] with various packages and scripts. Table 2.5-1 includes all the scripts and packages used in this project. The custom scripts written by me are available at. 'https://github.com/RGmetab/Thesis_Scripts'.

Table 2.5-1: List of packages and scripts used in statistical analyses.

Name	Function	References
cowplot	R package, extension for producing complex plots.	[194]
ggplot2	R package, for producing plots.	[195]
mixOmics	R package, set of tools for metabolomics analysis.	[196]
NMRMetab_norm_scale.R	In-house script, performs normalization and scaling for metabolomics data. Written by Dr Arturas Grauslys	
Pathway_Fisher.R	In-house script, performs Pathway analysis *via* Fisher's exact test. Written by me.	
PathTabforFisher.py	In-house script, generates pathway table from KEGG database to be used in Fisher's exact test. Written by me.	
pvca	R package, set of tools for variance estimation.	[145]
readBruker.R	Published script, reads in NMR spectrum in bruker format. Written by Dr Jie Hao.	[197]
reshape2	R package, arranges data frames for ggplot functions.	[198]
sourceARSyN.R	Published script, loads the required functions for ARSyN batch correction.	[142]
sva	R package, set of tools for batch correction.	[199]

2.5.1 Data processing

2.5.1.1 Data cleaning

Spectra that did not pass the QC (as described in section 2.3.6) were contrasted with lab book records to link the poor sample quality to known possible sources of variation. When suspected contamination, a comparison between spectral signals of the sample to spectral signals from the pooled data was done. Contamination signals were defined as peaks present in the question sample but not present in 99% of the total data. These spectra were deemed unsuitable to analyse and removed before analysis.

2.5.1.2 Data normalisation

Normalisation is an essential step in omics analysis and corrects fluctuations on sample preparation that result in a level of inter sample error. Probabilistic quotient normalization (PQN, [200]) is a widely used method in the NMR community and our chosen method. This method normalises each spectrum by a reference one, in this case we chose the median spectrum. These are the steps implemented in the PQN normalisation function:

1. Create a reference spectrum by calculating the median spectrum from the whole dataset (through bins).
2. To create the quotients, divide all spectra by the reference spectrum.
3. Calculate the median of the quotients to give the normalisation factor.
4. Divide the raw data by the normalisation factor.

2.5.1.3 Batch effect assessment and correction

The number of samples acquired in this project required the experimental design to include several batches for acquisition and preparation. All batches were designed such that a batch of any experiment contained all the experimental groups for even representation across batches. Under the assumption that batches will be inherently different due to experimental fluctuations, batch effect contribution was assessed using PVCA [145]. PVCA utilises PCA and variance component analysis in order to estimate the proportion of different factors in the dataset. Using PVCA the batch correction methods were also assessed in order to determine which method to use in case it was needed. Methods tested were; ASCA [ANOVA simultaneous component analysis] removal of systemic noise (ARSyN) [142], surrogate variable analysis (SVA) [199], and combining batches (ComBat) [143].

2.5.1.4 Data scaling and centring

Data was scaled to avoid magnitude-bias in variable selection when undertaken multivariate modelling. The chosen method was autoscaling. Autoscaling calculates the mean and standard deviation of each variable (bins). Then it subtracts the mean value from all individual points and divides by the standard deviation.

2.5.2 Multivariate Analyses

Univariate analyses test differences of one variable in a number of groups. Each test result has a probability of being a false positive or a false negative. Thus when multiple testing it is critical to perform false positive corrections. When looking at multivariate data having large number of variables to compare and not only wanting to know how if each individual one presents a significant difference between our groups of study but also whether there is an interaction between any of them and whether the overall fingerprint can be considered different from others. Furthermore, applying false positive correction methods in a very large number of tests usually results in an underpower analysis (see Figure 2.5-2). Thus, for multivariate data multivariate analysis are more suited. Below there is a description of the chosen multivariate approaches for this project.

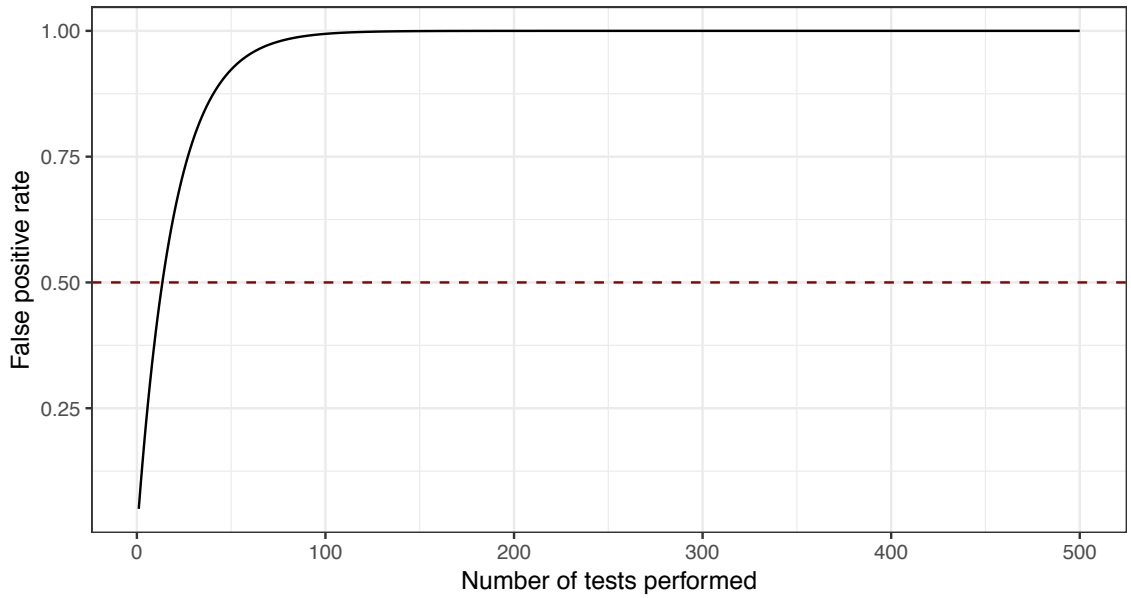


Figure 2.5-2: Increase of false positive rate during multiple hypothesis testing with significance level of α=0.05. Without p-value adjustment for multiple testing; at 14 tests false positive rate is higher than 50%, at 28 tests higher than 75% and at 45 tests higher than 90%.

2.5.2.1 Principal component analysis (PCA)

Principal component analysis (PCA) is an orthogonal data transformation that returns unobserved (latent) variables named principal components (PC). Each PC is the linear combination of the original variables in such a way that the first PC explains the most variance in the data. The second components explaining the most variance unexplained by the first PC and is orthogonal and uncorrelated to the first PC. Subsequent PCs follow the same procedure. This transformation results in a dataset where the original variables are replaced by uncorrelated PCs. In this new dataset of PCs only the first few (typically) PCs are required to explain majority of the variance in the dataset hence reducing the high number of starting variables with minimal information loss.

In this project PCA was preferred as a data exploration tool to reveal and observed the hidden structures in the data. on a PCA scores plot, each point represents a sample and can gives information about the similarities between samples. This is elucidated from the distances between points on the scores plot. Furthermore, we can link observed structures in the PCs to the original variables/metabolites. This is typically represented in loading plots although, selecting metabolites especially inherintly multicollinear data (e.g. NMR data) is often complicated.

2.5.2.2 Partial least square discriminant analysis (PLS-DA)

Partial least square - discriminant analysis (PLS-DA) is a variation of partial least square (PLS) regression. PLS is a supervised statistical model applied to multivariate datasets in order to make predictive models between two matrices. A PLS model requires a matrix of input data (predictors) and a secondary matrix called (response) where the output of the model is recorded. PLS projects both of these matrices (predictors and response) into two new matrices where the covariance between the two are minimised. PLS and PCA show some similarity in terms of creating latent (unobserved) variables, using of projections to new spaces although. While PCA projects maximum variance in latent variables called PCs, PLS projects predicted and observable variables using a linear regression model in latent variables called variates. PLS-DA variation uses a nominal vector as response which allows the building of models for classification problems. PLS models are particularly suitable for data with multicollinearity in predictors such as, NMR data where a single metabolite can be represented by multiple peaks depending on its molecular structure.

One critical step of statistical modelling is to check for overfitting and underfitting. Supervised statistical models requires the model to be trained on data prior to prediction. Models depending on creation of latent variables (not exclusively) such as PLS-DA can be heavily affected by this process. When the number of latent variables to be used in model set very high, usually to achieve high accuracy. By doing so, the model starts to train on how to recognise the noise in the data that is very specific to the training dataset. This phenomenon is also known as overfitting, as can be easily spotted by making predictions on new data. Since, experimental noise is random, an overfitted model will have poor prediction accuracy on new data. On the opposite side of the scale the models can be underfitted. This is when too few latent variables a chosen and the model does not perform well due to not being able to describe the data in selected number of latent variables. This also results in poor prediction accuracy. In both cases any further analysis (e.g. variable selection) would be unreliable. This can be avoided by tracing the training of the model by making prediction on a new dataset which the model is never exposed to during training. The ideal model building process would require 3 dataset one for training, another for validation of the trained model and a final one for testing. Although this is the best way to build a model, this is not always possible (especially in metabolomics studies). For such cases a method called cross-validation can be used. In cross-validation, data is randomly split into training set and test set (typically 90% to 10% respectively). Following the split the model is trained on training set followed by prediction on the test set. After prediction, the model is typically assessed by prediction

accuracy. This process is repeated, multiple times (between 100-150 is a good rule of thumb) and for different numbers of latent variables. Typically model with the best predictive accuracy is selected.

In this project PLS-DA was used to build predictive models between experimental groups. PLS-DA model's performance was assessed by dividing the data into 70% and 30% training and validation sets respectively. The training set was then 5-fold cross-validated over 250 repetitions. Classification errors were used to determine the optimal model complexity parameter. The refined model was then used to predict the validation set to obtain model performance. Model performance was assessed by calculating accuracy, precision, recall and F-score as describe in Sokolova and Lapalme [201]. R^2 and Q^2 score were not preferred in model performance assessment since they require the response matrix to be on a ratio scale rather than nominal. Using the refined model parameters, a model was built using both training and validation sets, yielding a generalised model of the dataset. These final models were then used in bin (representing metabolites) selection to reveal key metabolites driving the differences between experimental groups.

2.5.2.3 Bin selection

2.5.2.3.1 Variable importance of the projection (VIP)

Bin (variable) selection from a statistical model is critical step for most metabolomics studies in order to extract biologically relevant information. PLS-DA and its derivative methods are designed to transform the data and make predictions. However, variable (bin) selection is not integrated into the model building process. Variable importance of the projection (VIP) scores is such a method that is often preferred with datasets with multicollinearity. VIP scores in essence are weighted sum of squares of PLS weights (calculated during PLS-DA model building) which also take explained variance in PLS variates. This method is designed to be used for multivariate datasets where there is correlation between variables as well as a higher number of variables than samples.

Once the VIP scores are calculated a cutoff needs to be defined in order to include or exclude variables. VIP scores are calculated as such that average of all VIP scores squared is 1 [202]. Hence a cutoff of 1 was used to select important variables. Due to the nature of VIP scores calculation, after excluding variables via VIP scores, a refined PLS-DA model built after will have VIP scores less than 1 for previously passing variables.

Due to the nature of NMR data, metabolites with multiple signals presented multiple entries in VIP scores. We then included an extra step to select the most representative bin per metabolite to take forward on the analysis pipeline.

2.5.2.3.2 Correlation Reliability Score (CRS)

To address the problem of selecting appropriate representative bins from VIP scores a correlation-based scoring method was developed. Depending on the molecular structure some metabolites have multiple NMR signals. These multiple signals arising from a metabolite will increase/decrease as the concentration increases/decreases, yielding a high correlation score between multiple signals belonging to the same metabolite. However, some areas of the spectra are populated by peaks belonging to multiple metabolites, thus some bins might be more reliable markers for a metabolite than other. Correlated bins (assigned to the same metabolite) were scored to determine their reliability to report on the assigned metabolite, a specific challenge given the severe overlap of peaks in certain bins of the NMR datasets. This score is referred to as the correlation reliability score (CRS) that shows the performance of a bin as a representative of the metabolite through its correlation with other bins of the same metabolite. This score is determined via the following algorithm:

1. Calculate Pearson correlation matrix for all the identified bins per metabolite.
2. For each unique metabolite extract individual bin correlation values.
3. Calculate the mean for each individual bin of the unique metabolite.
4. Multiply each score by 100 to present the percentage.

It should be noted that algorithm above is designed to select a representative bin from a metabolite with multiple NMR signals. Hence it is not suitable for metabolites with single NMR peak. When the algorithm is used on metabolites with single NMR peaks a score of 100% would be calculated due to self-correlation. When tabulated these signals were noted as not applicable (NA). In order to separate the candidate bins from non-candidate bins, a passing score was then calculated using the previously calculated CRS scores via the following algorithm:

1. Exclude all bins with a 100% CRS.
2. Calculate median and standard deviation with the remaining scores.
3. CRS_{pass} = median – standard deviation

A CRS above the threshold represents a high correlation of the bin to the rest of the signals of the same metabolite. For final selection, highest CRS scores of non-overlapping bins (where applicable) were preferred yielding a dataset consisting of one variable per metabolite to be used in univariate and pathway analyses. Prior to further analysis the efficacy of the selection was assess by PCA scores plot and PLS-DA model predictions with the dataset of selected metabolites.

2.5.3 Univariate analyses

Univariate analyses were performed using Welch's t-test and analysis of variance (ANOVA) where appropriate in order to compare the means of selected metabolites. To account for type-I errors arisen from multiple hypothesis testing, P-values were corrected *via* Benjamini & Hochberg (BH) [203] method unless otherwise stated. To visualise the significant changes in metabolites between experimental groups, boxplots (Figure 2.5-3) were plotted.

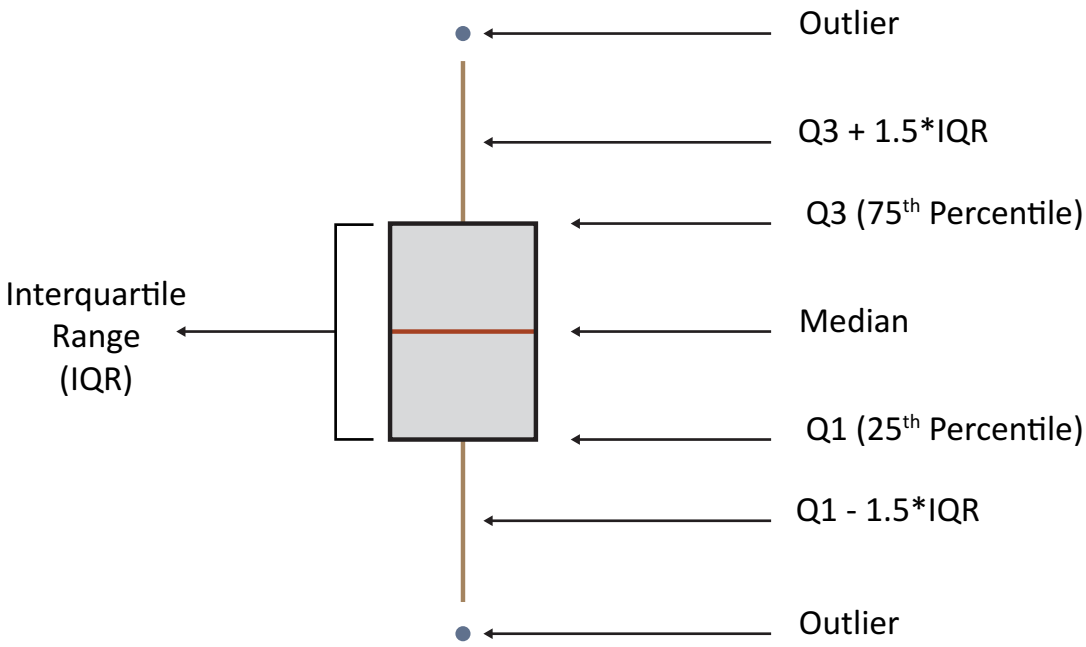


Figure 2.5-3: Break down of a boxplot. Outlier limits were calculated as Q1 - 1.5*IQR for lower and Q3 + 1.5*IQR for higher, where IQR is Q3-Q1.

2.5.3.1 Welch's t-test

Following metabolite selection, for further insight metabolites were compared between experimental groups. Welch's t-test was used to compared the means where only 2 experimental groups were present. Welch's t-test [204] is an adaptation of Student's t-test [205] that does not assume equal variance between groups. Welch's t-test operates under

the null hypothesis where there is no difference between the means of the groups being tested. A t-score is calculated and compared against a t-distribution probability table for analysis. The p-value shows the probability of obtaining such results in the following tests given that the null hypothesis is true. An arbitrary significance level of α=0.05 was used to either accept or reject the null hypothesis. It should be noted when conducting multiple t-tests, the chance of finding false positives increase. To avoid drawing conclusions from false positives, p-values were adjusted with BH method for false discovery rate (FDR).

2.5.3.2 Analysis of variance (ANOVA) & Tukey's honest significant difference (HSD)

ANOVA was used where Welch's t-test was not applicable due to the number of groups present. ANOVA compares the means of multiple groups while operating under the null hypothesis that there is no difference between the means of the groups being tested. Determination of significance was done by the assessment of p-values of an arbitrary significance level of α=0.05. ANOVA does not perform pairwise comparisons; results reported to be significant mean there is at least one group that is different from at least one other group. Pairwise comparisons need to be performed by a post-hoc test such as Tukey's honest significance difference (HSD).

Tukey's HSD is a pairwise comparison, where means of all the groups are compared to every other group. The difference is deemed significant via p-value with the widely adopted significance level of α=0.05. Tukey's HSD performs p-value adjustment for multiple comparisons built into the comparison.

2.5.4 Metabolite set enrichment analysis (MSEA) and interpretation

Upon the selection of metabolites, through PLS-DA modelling a qualitative metabolic set enrichment analysis (MSEA) was used based on a Fisher's exact test. MSEA provides a probability measure for a set of metabolites likelihood of representing a pathway in a system. Given both the qualitative nature of this analysis and metabolomics showing 'metabolic snapshot', it is not possible to annotate a pathway as being increased/decreased or up-regulated/down-regulated. MSEA's sole purpose is to provide possible leads on pathways which to be explored and discussed further in the light of complimentary data and/or other publications.

In this project pathway analysis was performed using a number of metabolite sets (one per pathway) generated from the KEGG database (accessed on 26/03/2019). The metabolite set for *An. gambiae* and *Ae. aegypti* were curated from KEGG [206] using a custom scripts written by me (available at 'https://github.com/RGmetab/Thesis_Scripts'). The curated *An. gambiae* metabolite set included 4270 (of which 2770 were non-repeating) metabolites and 135 pathways. The *Ae. aegypti* metabolite set comprised 129 pathways with 3991 metabolites of which 2697 were non-repeating. Identified metabolite names were converted to KEGG compound codes which were then used to calculate the probability of individual pathways via a one-sided Fisher's exact test with Expression Analysis Systematic Explorer (EASE) scoring as applied in DAVID [207]. EASE applies more stringent rules on this calculation by using a penalty (number of hits - 1) on the calculation. This penalty also ensures any pathway found with only one metabolite is excluded. Resulting p-values were adjusted for Type I errors with BH, with pathways with p-values less than 0.05 presented as significant and discussed further in the relevant results chapters.

Chapter 3

3 Investigation of sex-specific metabolic differences in mosquito species of *An. gambiae* knock-down (Cyp4g16 & Cyp4g17), wild type *An. gambiae* and wild type *Ae. aegypti*

3.1 Introduction, chapter aims and objectives

In mosquito research, sex has always been a key factor in dictating the research focus. A great deal of mosquito research is on female mosquitoes due to them being the vector capable of pathogen transmission with great efficiency. Unless there is a particular interest in males, they are discarded. This extensive focus on females has generated a larger amount of knowledge for one sex than the other. Currently, there are a wide range of publications and books about the anatomical differences between males and females. Unfortunately, this is not the case for sex-related metabolic differences. Further exploration of these differences can shape how non-female specific experiments (e.g. blood meals and oviposition) are designed.

In the mosquito life cycle from eggs to pupae, males and females do not show many morphological differences. They behave and act very similar. Even in the pupal stage where the most morphological differences are observed, it is only possible to distinguish between males and females under a microscope. A study investigating diapause in moths (Antheraea mylitta) performed a comparison between males and females fed on different crops and found that the amino acid (except one crop) and protein content were not significantly different [208]. Although this study was on moth's, metamorphosis is relatively well conserved in insects and so can be expected to be similar in mosquitoes [209]. Thus, strong differences in the metabolic profiles of males and females are not expected. On the other hand, female adults are required to find a blood source for oviposition, which may require the mosquito to maintain flight for a longer time. Therefore, a metabolic profile comparison of energy storage mechanisms and related metabolites would be expected to be higher in females compared to males.

This chapter aims to explore whether metabolic sex-specific differences can be detected using NMR metabolomics and if so, what effect sex has on the metabolic profile of pupa and adult mosquitoes across species and strains. These points will be explored under three main male and female comparisons.

1) Cyp4g16 and Cyp4g17 knockdowns of An. gambiae
2) 2) Wild type *An. gambiae*
3) 3) Wild type *Ae. aegypti.*

To achieve this, PCA scores plots of metabolite bins will be used to determine the major contributors in variance. If sex is not one of these major contributors, a statistical modelling approach will be applied to determine to what extent sex differences are observable via polar NMR metabolomics.

It is known that mosquitoes use certain hydrocarbons (HC) for communication purposes, such as pheromones, multiple branched alkanes and alkenes. Cuticular hydrocarbons (CHC) are utilised for waterproofing and physical protection properties, such as alkanes and single branched alkanes and they are not known to differ in variety between males and females [78], [86]. The biosynthesis of CHCs mainly takes place in the pupal stage and early adult stage. CHC biosynthesis is a culmination of several complex pathways and a wide variety of final products can be synthesised [78], [172]. The majority of CHCs can be grouped into families

The enzyme Cyp4g16 has been shown in vitro to catalyse the decarbonylation of such CHCs and the enzyme Cyp4g17 has been hypothesised to also carry out this reaction [75]. Performing a metabolomics analysis of such knock-downs to investigate the changes in the precursors without exploring the difference between males and females may lead to false conclusions confounding with sex differences. Prior to investigation of CHC biosynthesis precursors, a comparison between males and females was first carried out.

3.2 Experimental Design

All samples used in this project were field collected and are generations of the same lineage. All mosquito breeding was done in the Liverpool School of Tropical Medicine insectary. Table 3.2-1 shows the history of the species used in this project.

Table 3.2-1: Metadata of samples used in this project.

Species	Genotype	Pyrethroid	Strain	Origin	Since
An. gambiae	Knock-down	Susceptible	Gal4	Lab	2011
		Susceptible	RNAi16	Lab	2014
		Susceptible	RNAi17	Lab	2014
	Wild type	Resistant	VK7	Burkina Faso	>5 years
		Susceptible	N'Gusso	Cameroon	>5 years
Ae. aegypti	Wild type	Resistant	Cayman	Cayman Islands	>8 years
		Susceptible	New Orleans	USA	>8 years

In order to establish whether sex influenced metabolic profiles in each strain, samples were grouped as male and female. All samples were collected as biological triplicates and sample numbers post QC are shown in Table 3.2-2.

Table 3.2-2: Sample numbers for sex testing post QC. Criteria used for QC were: overall flat baseline, TSP HHFW ≤ 1 Hz, water residue ≤ 0.25 ppm and no presence of contaminants.

Samples collected

	***An. gambiae* knock-down**						***An. gambiae* wild type**				***Ae. aegypti* wild type**			
	KD16		**KD17**		**Control**		**Susceptible**		**Resistant**		**Susceptible**		**Resistant**	
Gender	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀
Pupa	15	15	15	15	15	15	15	15	15	15	15	15	15	15
Adult	15	15	15	15	30	30	15	15	15	15	15	15	15	15

Spectra acquired

	***An. gambiae* knock-down**						***An. gambiae* wild type**				***Ae. aegypti* wild type**			
	KD16		**KD17**		**Control**		**Susceptible**		**Resistant**		**Susceptible**		**Resistant**	
Gender	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀
Pupa	7	10	9	11	10	10	15	15	15	15	15	15	15	15
Adult	15	15	15	15	30	30	15	15	15	15	15	15	15	15

Quality control passed

	***An. gambiae* knock-down**						***An. gambiae* wild type**				***Ae. aegypti* wild type**			
	KD16		**KD17**		**Control**		**Susceptible**		**Resistant**		**Susceptible**		**Resistant**	
Gender	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀
Pupa	4	7	7	8	5	7	12	15	11	9	15	12	14	14
Adult	12	15	14	15	28	28	13	14	15	14	9	11	10	11

♂: Male; ♀: Female; **KD16**: Cyp4g16 knock-down; **KD17**: Cyp4g17 knock-down

3.3 Spectral binning and metabolite identification

For statistical analysis, all peaks were manually selected to create a list of chemical shift regions. These regions were collated in a pattern file. Metabolites were assigned using a combination of 1D & 2D in-house libraries and Chenomx software by matching chemical

shifts, intensity ratios between groups of peaks and peak multiplicity. Assigned metabolites were then associated with their specific chemical shifts in the pattern file created (Appendix 1). Representative 2D assignments for *An.gambiae* knock-down, *An. gambiae* wild type and *Ae. aegypti* wild type can be found in Appendix 2, Appendix 3 and Appendix 4 respectively. Assigned pattern files were verified by highlighting the chemical shift regions on a representative spectra (Figure 3.3-1 & Figure 3.3-2). A total of 496 chemical shift regions were selected in the binning process. For *An. gambiae* data, out of 496 bins 107 (21.57%) were associated with 21 metabolites. For *Ae. aegypti* data, 111 bins out of 517 (21.47%) were associated with 21 metabolites. Throughout the statistical analyses, unidentified bins were included until metabolite selection. In all data sets, the annotated metabolites were identical and are shown in

Table 3.3-1. The annotation table shows the list of assigned metabolites with their confidence levels according to MSI, corresponding KEGG compound ID, and their compound classification.

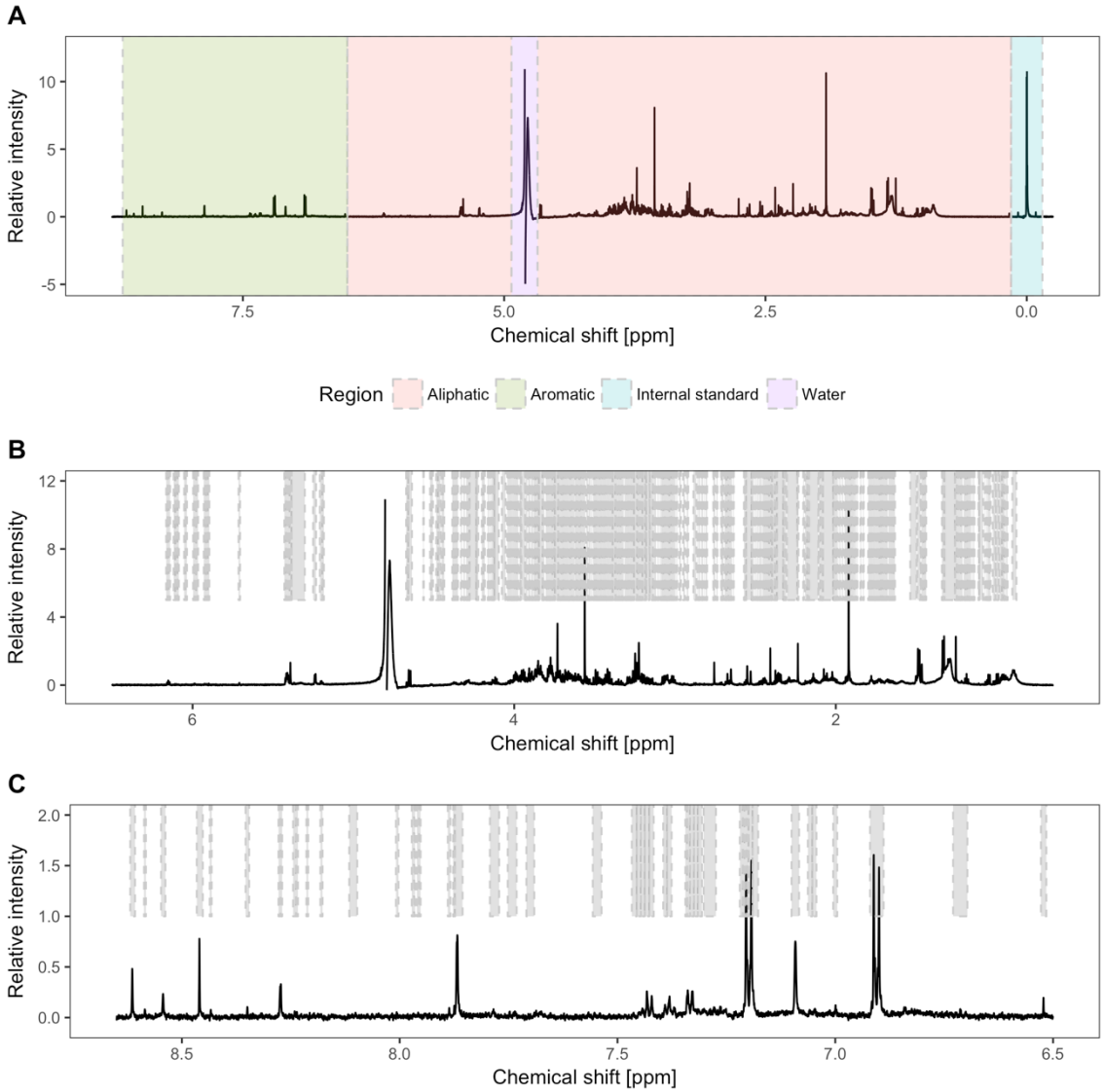


Figure 3.3-1: NMR spectrum of an *An. gambiae* knock-down pupa control. The y-axis represents peak intensity relative to water peak (purple). A) NMR spectrum highlighted by typical regions of interest. Spectrum shown is between 0-9 ppm where signals were observed. Only the water residue and TSP signals were truncated. During the binning process, a total of 496 bins were selected. B) Aliphatic region of NMR spectrum typically between 0-6.5 ppm, total number of bins in this region is 447. C) Aromatic region of NMR spectrum typically between 6.5-9 ppm, total number of bins in this region is 49. The grey boxes in B and C represent manually selected bins.

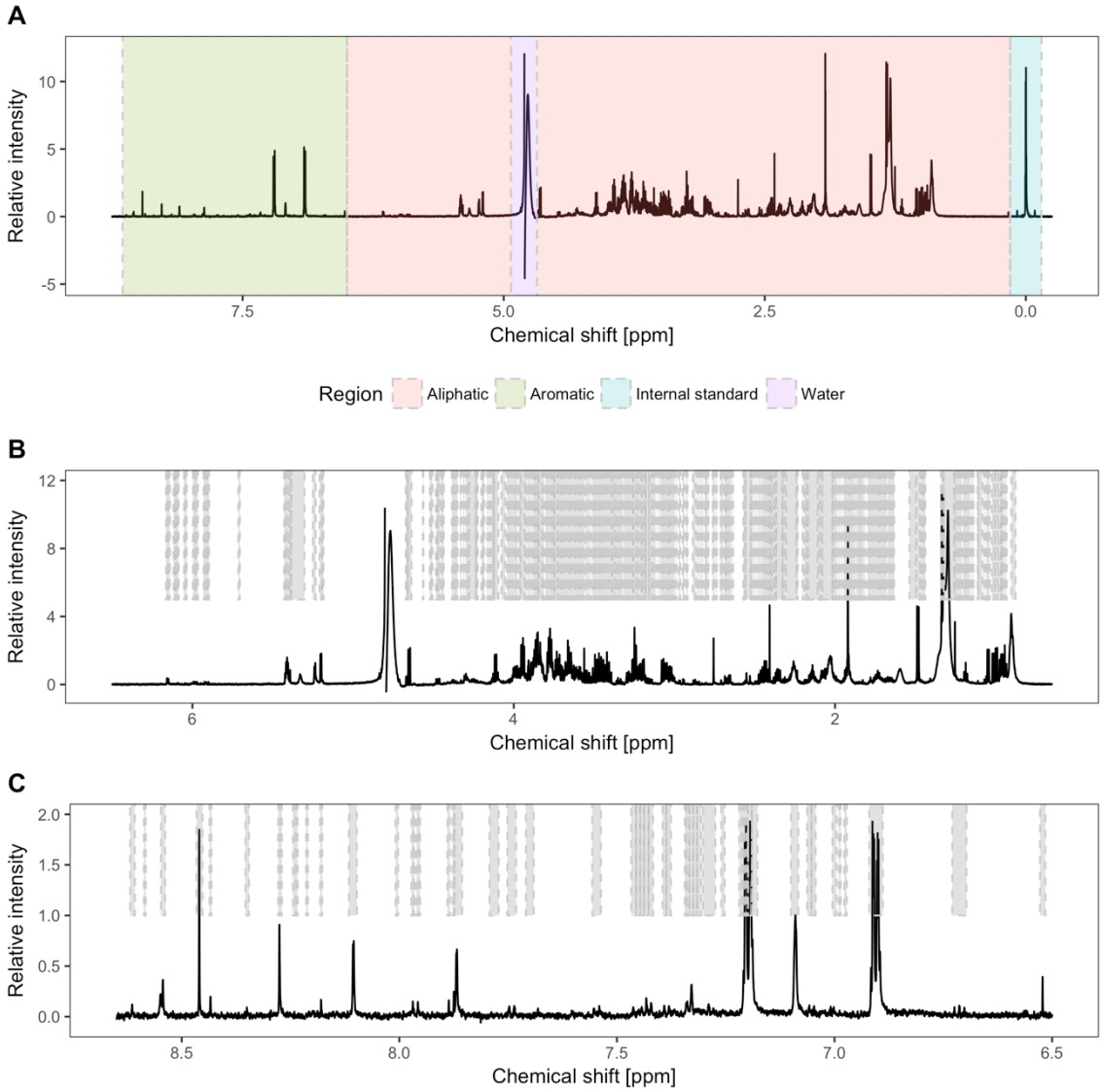


Figure 3.3-2: NMR spectrum of an *Ae. aegypti* pupa. The y-axis represents peak intensity relative to water peak (purple). A) NMR spectrum highlighted by typical regions of interest. Spectrum shown is between 0-9 ppm where signals were observed. Only the water residue and TSP signals were truncated. During the binning process a total of 513 bins were selected. B) Aliphatic region of NMR spectrum typically between 0-6.5 ppm, total number of bins in this region is 461. C) Aromatic region of NMR spectrum typically between 6.5-9 ppm, total number of bins in this region is 52. In figures B and C, grey boxes represent manually selected bins.

Table 3.3-1: Metabolite assignment table, with MSI level [191], KEGG compound code and classifications. See Appendix 1 for full assignment table.

Classification	Metabolite	Metabolite Identification Level (MSI)	Bins			KEGG code
			Unique	Overlap	Total	
Alcohols	Methanol	Level 1	1	0	1	C00132
Amino acids	Alanine	Level 1	2	3	5	C00041
	Glutamate	Level 1	7	3	10	C00025
	Glutamine	Level 1	2	3	5	C00064
	Glycine	Level 1	1	0	1	C00037
	Isoleucine	Level 1	3	0	3	C00407
	Threonine	Level 1	4	0	4	C00188
	Tryptophan	Level 1	12	3	15	C00078
	Tyrosine	Level 1	12	2	14	C00082
	Valine	Level 1	5	0	5	C00183
Carboxylic acids	Acetate	Level 1	1	0	1	C00033
	Formate	Level 2a	1	0	1	C00058
	Fumarate	Level 2a	1	0	1	C00122
	Lactate	Level 1	4	0	4	C00186
	Propionate	Level 2b	6	0	6	C00163
	Pyruvate	Level 1	1	0	1	C00022
	Succinate	Level 1	1	0	1	C00042
Purines	Oxypurinol	Level 2b	1	0	1	C07599
	Xanthine	Level 2b	1	0	1	C00385
Saccharides	Glucose	Level 1	24	8	32	C00031
	Trehalose	Level 1	10	7	17	C01083

Representative raw ^{1}H-NMR spectra showing male and female comparisons that can be found in Appendices are: *An. gambiae* knock-down pupae (Appendix 5), *An. gambiae* knock-down adult (Appendix 6), *An. gambiae* wild type pupae (Appendix 7), *An. gambiae* wild type adult (Appendix 8), *Ae. aegypti* wild type pupae (Appendix 9) and *Ae. aegypti* wild type adult (Appendix 10).

3.4 Sex-specific differences in metabolic profiles of *An. gambiae* knock-down (Cyp4g16 & Cyp4g17) pupa and adult

3.4.1 Sex differences in mosquito pupae

3.4.1.1 Statistical analysis

Prior to statistical analysis, the contribution of batch effect was estimated using PVCA. Experiments of this scale require the experimentation to be performed in batches. Although most environmental factors are controlled, such as temperature, day and night cycle, humidity and feeding, variation between batches can still be observable. These variations can arise due to other factors such as activity of sample in their environment, their feeding habits, and their natural developmental synchronicity, which can cause unwanted clustering

in the data. Sometimes, this clustering behaviour in data can interfere and/or mask the signals arising from experimental effects. In order to prevent such an effect, variance in the data can be estimated and batch correction methods can be applied if necessary. Figure 3.4-1 shows the high batch effects, present in the raw pupa data. In the raw data (red bars), batch effects were found to be accounting for 81.20% of the variance in the data. Upon applying different correction methods, ComBat (blue bars) was found to be the most effective in reducing the batch variation to 1.42%, and was used in all subsequent analyses.

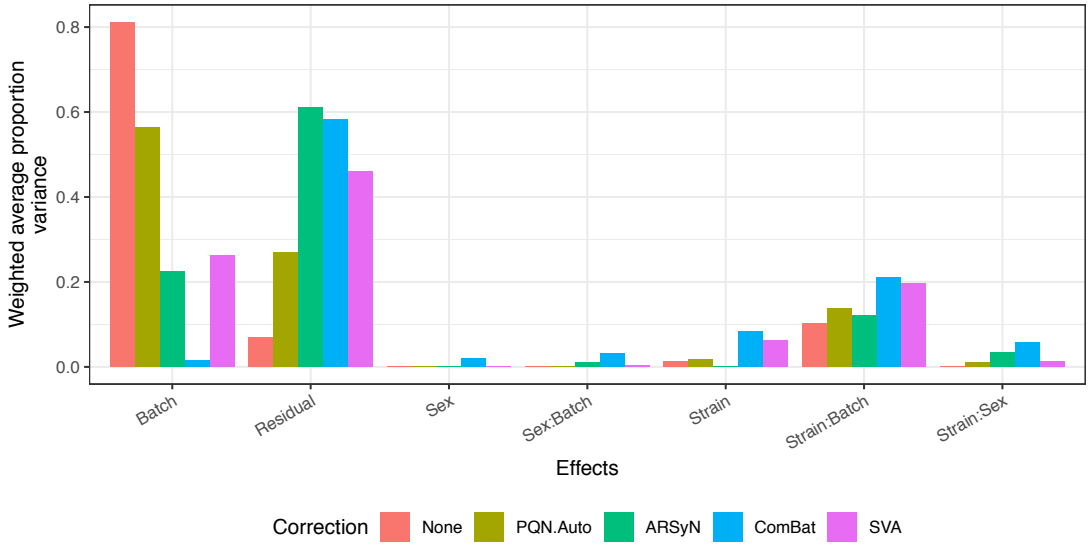


Figure 3.4-1: Estimation of batch, sex and strain variances *via* PVCA in knock-down *An. gambiae* pupae. Colons in the x-axis denote the combined effects of two sources of variance and, Residual denotes all remaining variation in the data (residual). ComBat outperformed other methods with the lowest batch variance. Correction methods applied: None, no correction; PQN.Auto, probabilistic quotient normalization and autoscaling; ARSyN, ASCA [ANOVA simultaneous component analysis] removal of systematic noise; ComBat, Combining batches and SVA, surrogate variable analysis.

It is important to establish the major variances in the data before further analysis. Sex in these samples is one of the major concerns that may affect the results. To resolve this, PCA (Figure 3.4-2-A) was used to observe the sources of the major variance in the data. The plot displays PC1 (18.43%) against PC2 (12.34%), with no clear distinction between male and female species. The first two components account for 30.77% of the explained variance in the metabolic profiles and a total of 25 components were required to explain the 95% variance in the data. This is consistent with the theory that the overall metabolic profile of pupa knock-downs is not majorly influenced by sex.

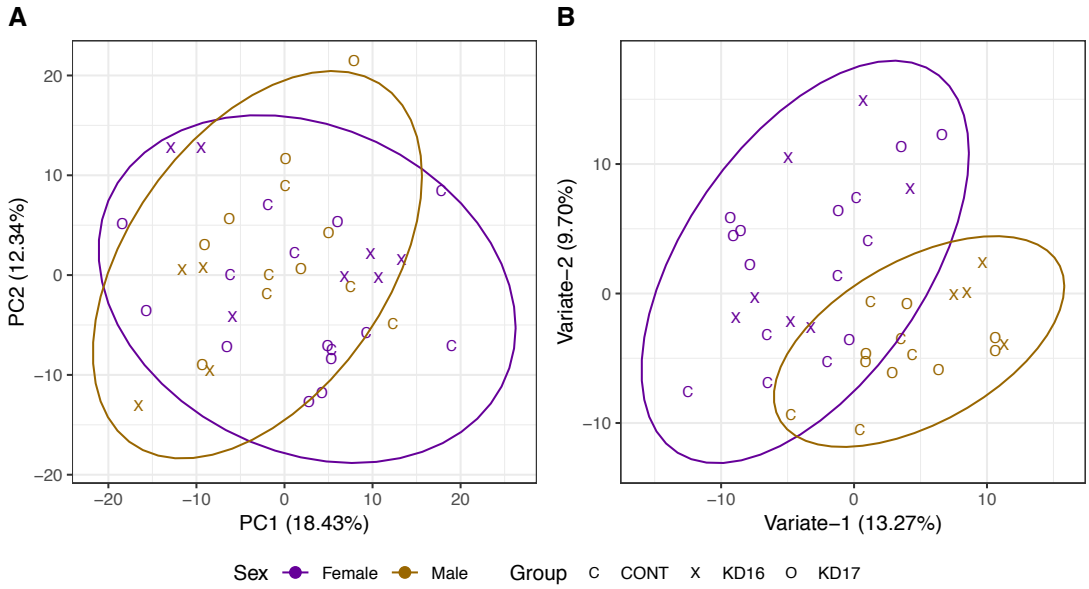


Figure 3.4-2: A) PCA scores of all knock-down *An. gambiae* pupae coloured by sex (n_{male} = 16 and n_{female} = 22). Brackets report the variance explained by the PC. Twenty-five PCs were required to achieve 95% explained variance. Only PC1 (18.43%) and PC2 (12.34%) are shown in the Figure for simplicity. Clear clustering according to sex is not observed suggesting the metabolic differences between males and females very high variations in the dataset. Sex difference was not observed in the density plots of the other PCs. Ellipses represent 95% confidence region. B) PLS-DA scores of pupae knock-down of *An. gambiae* discriminated by sex (n_{male} = 16 and n_{female} = 22). Model complexity of three variates was optimised *via* cross-validation with 45.45% accuracy. Two clusters can be easily seen in the scores plot indicating the discrimination capabilities of the model. Ellipses represent 95% confidence region.

In order to probe deeper, a statistical model can help to bring out the sex effects masked by other components of the data. A cross-validated PLS-DA model was performed on the same dataset to establish whether any metabolic differences could be attributed to the sex differences in pupa. A three-variate PLS-DA model (Figure 3.4-2-B) was determined to be optimal to discriminate between females and males with 45.45% accuracy (Appendix 11 for further metrics) discriminating between females and males. Nearly half the sample were correctly classified by the model. Using VIP scoring as a criterion, metabolites influential in such discrimination can be extracted.

3.4.1.2 Key metabolites between males and females

VIP scores were calculated from the model in order to select spectral features influencing the separation between males and females. A passing one was applied to VIP scores of variate-1 and variate-2 where separation between males and females was observed. A total of 103 bins scored above the passing score where only 24 (23.30%) were identified. Figure 3.4-3 shows the 24 identified bins representing 10 unique metabolites. In order to assess their representative qualities, CRS was applied.

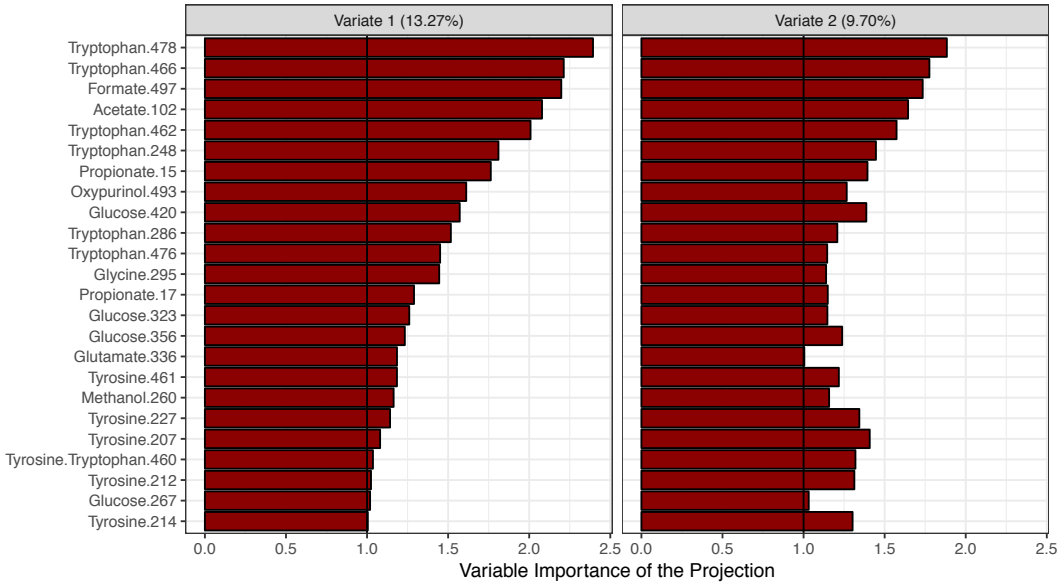


Figure 3.4-3: VIP scores of PLS-DA model built on sex discrimination of KD *An. gambiae* pupae. Identified bins scoring higher than 1 on variate-1 and variate-2 were selected from the PLS-DA model. Black line represents VIP score of 1.

As a secondary criterion, CRS was used to select representative bins for a metabolite. CRS (Table 3.4-1) was calculated for all identified bins and filtered to show bins selected by VIP scoring. CRS threshold score of 27.11% was calculated using all the CRS scores.

Table 3.4-1: CRS scores of VIP selected bins from the knock-down *An. gambiae* pupa PLS-DA model for sex discrimination. Rep: representative bin.

Metabolite	Bin	CRS [%]	CRS > 27.11%	Rep	Metabolite	Bin	CRS [%]	CRS > 27.11%	Rep
Acetate	102	Singlet	NA	102	Tryptophan	476	49.88	✓	476
Formate	497	Singlet	NA	497		478	47.22	✓	
Glucose	267	64.41	✓	267		462	44.70	✓	
	356	58.19	✓			466	38.40	✓	
	420	57.84	✓			248	37.52	✓	
	323	54.35	✓			460*	32.65	✓	
Glutamate	336	-5.75	×	-		286	31.08	✓	
Glycine	295	Singlet	NA	295	Tyrosine	214	97.76	✓	214
Methanol	260	Singlet	NA	260		460*	97.64	✓	
Oxypurinol	493	Singlet	NA	493		212	97.49	✓	
Propionate	17	39.48	✓	17		227	97.03	✓	
	15	36.46	✓			461	94.27	✓	
						207	92.61	✓	

*, overlapping peak; NA, correlation score for singlet were not calculated.

Only non-overlapping (where applicable) bins scoring higher than the CRS threshold were considered as a representative of the metabolite. Table 3.4-2 shows the list of selected

metabolites and their representative bins, chemical shift of the bin and the metabolite's KEGG code. From the 10 metabolites identified *via* VIP scores, only nine were selected from CRS. Glutamate was excluded due to the low CRS. CRS selected metabolites comprised of five metabolite classes; alcohols, amino acids, carboxylic acids, purines and saccharides.

Table 3.4-2: List of selected metabolites *via* CRS and their representative bins.

Class	Metabolite	Bin	Chemical shift [ppm]	KEGG code
Alcohols	Methanol	260	3.36	C00132
Amino acids	Glycine	295	3.57	C00037
	Tryptophan	476	7.55	C00078
	Tyrosine	214	3.08	C00082
Carboxylic acids	Acetate	102	1.92	C00033
	Formate	497	8.46	C00058
	Propionate	17	1.07	C00163
Purines	Oxypurinol	493	8.27	C07599
Saccharides	Glucose	267	3.41	C00031

The metabolites selected were used in PCA and PLS-DA to verify their discriminatory properties exclusively. The PCA scores (Figure 3.4-4-A) plot of PC1 (44.78%) against PC2 (16.14%) exhibits a separation of the majority of the data points representing 60.92% variance in the data. In Figure 3.4-4-A four female samples on the left (two KD16 and two KD17), and four male samples (two CONT and two KD 17) did not cluster with their groups.

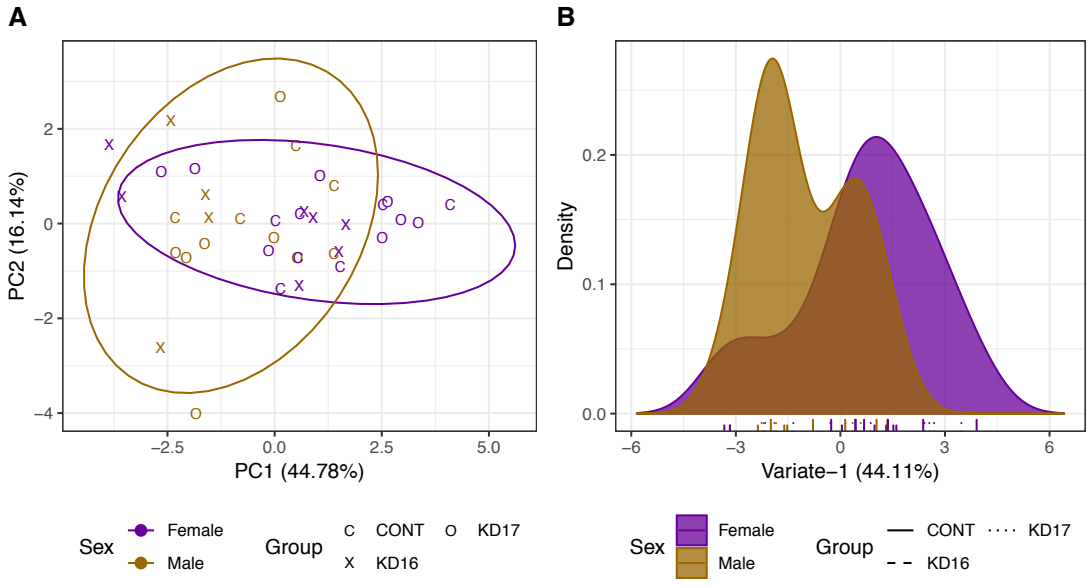


Figure 3.4-4: A) PCA scores of selected metabolites of *An. gambiae* pupae coloured by sex (n_{Male} = 16 and n_{Female} = 22). Clustering of sex can be observed along PC1 (44.78%). PC2 (16.14%) does not explain the sex difference in the metabolic profiles. A total of 25 components were required to achieve 95% explained variance. Ellipses represent 95% confidence region. B) PLS-DA density plot to verify metabolite selection in pupa knock-down of *An.* gambiae discriminated by sex (n_{Male} = 16 and n_{Female} = 22). Model complexity of one variate (44.11% explained variance) was optimised using a cross-validation with 63.64% accuracy. Ticks under the density plot represents samples from each group.

To assess sex discriminatory performance of the selected metabolites, a PLS-DA model was built using their representative bins. A single-variate PLS-DA (Figure 3.4-4-B) model was built with 63.65% predictive accuracy (Appendix 11 for further metrics) on unseen data at discriminating between males and females.

In order to gain metabolite level understanding on sex differences, metabolite levels were compared *via* BH adjusted t-test (for detailed statistics see Appendix 12) and shown by boxplots (Figure 3.4-5). A total of nine metabolites were selected from the PLS-DA model, although none were found to be significantly different in BH adjusted t-tests when considered in isolation. In the boxplots of three metabolites tryptophan, oxypurinol, and format sub-populations can be seen. These sub-populations are not consistent across all metabolites. Furthermore, the sample within the sub-populations do not follow a trend. Within these sub-populations clustering of KD17 & CONT, KD16 &KD17, and all three can be seen. These sub-population could not be attributed to any known metadata.

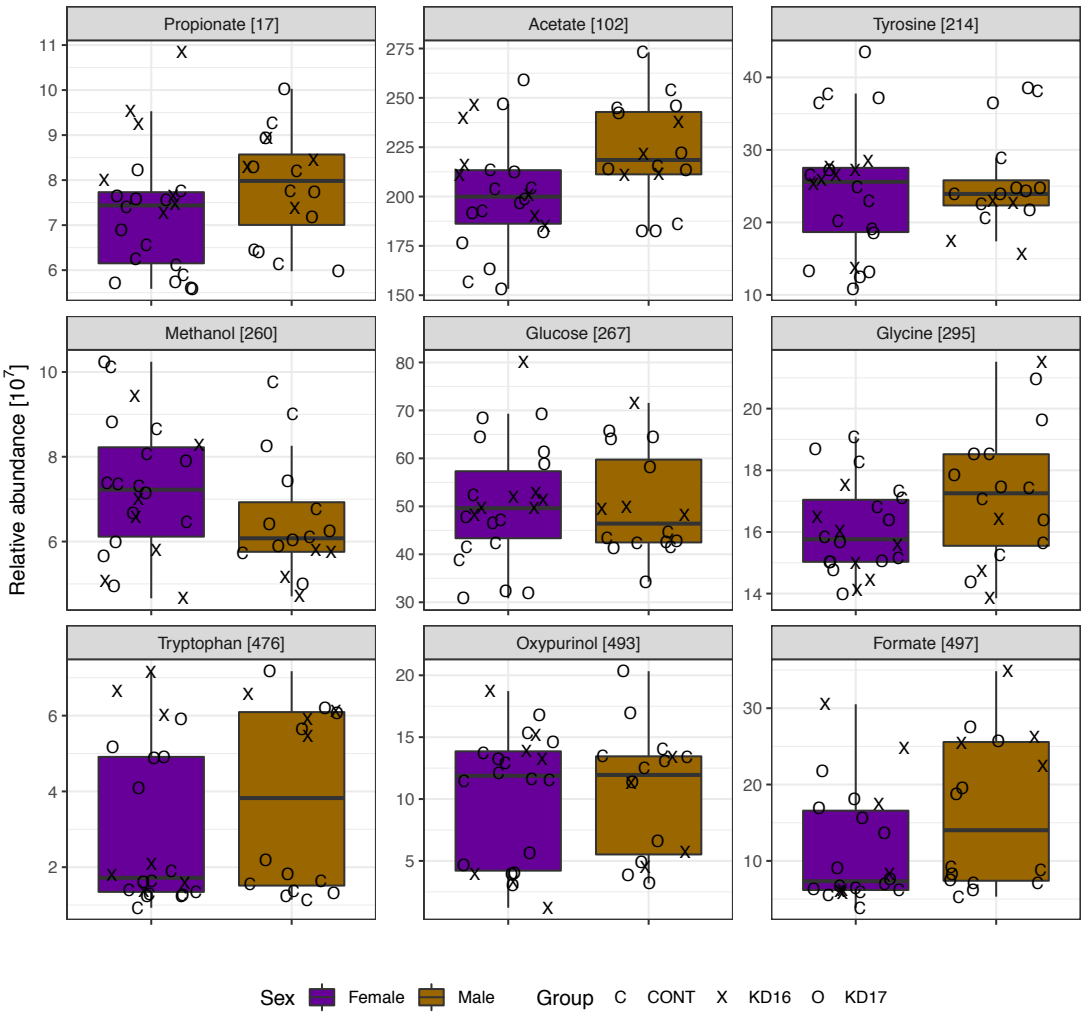


Figure 3.4-5: Boxplots of selected metabolites from *An. gambiae* pupae PLS-DA model on sex discrimination (n_{Male} = 16 and n_{Female} = 22). No significance differences were observed.

3.4.2 Sex-specific differences in adult mosquitoes

3.4.2.1 Statistical analysis

The same approach as was performed on pupae was taken to assess batch effect and the major source variance in the adult samples for the knock-down *An. gambiae*. PVCA plot (Figure 3.4-6) shows how much batch effect accounts for the variation in adult knock-down *An. gambiae*. Raw data (red bar) showed batch to be accounting for 00.03% in the data hence no batch correction was applied to the data. Through PVCA, it was also found that when PQN normalisation and auto-scaling (yellow bar) were applied, the proportion of the batch variation in the data reduced to 2.79×10^{-17}.

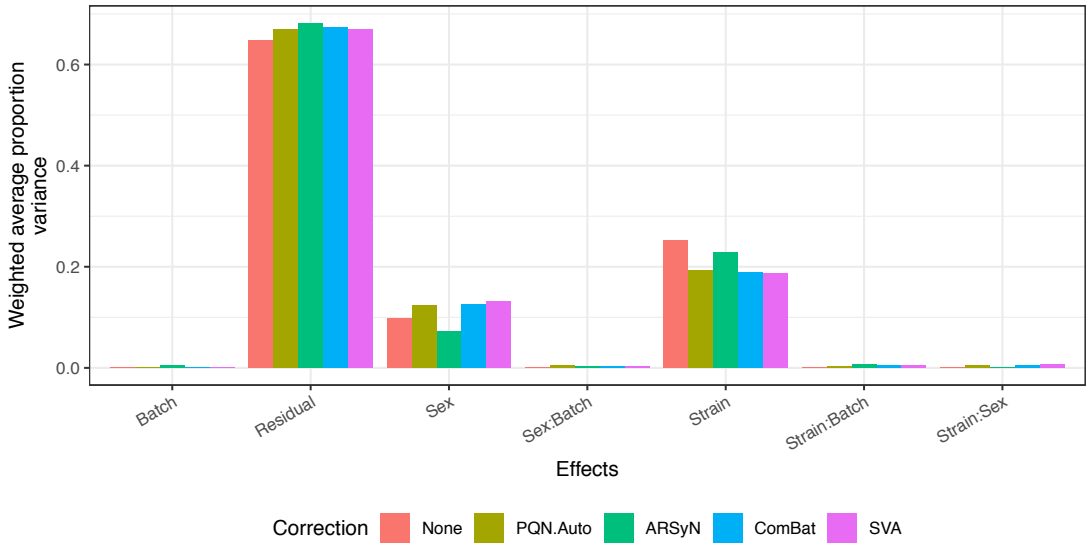


Figure 3.4-6: Estimation of batch, sex and strain variance *via* PVCA in knock-down *An. gambiae* adults. Colon denotes the combined effects of two sources of variance and Residual denotes all the remaining variation in the dataset (residual). PVCA shows batch effect is minimal and masked by the effects caused by sex and strain. Correction methods applied: None, no correction; PQN.Auto, probabilistic quotient normalization and autoscaling; ARSyN, ASCA [ANOVA simultaneous component analysis] removal of systematic noise; ComBat, Combining batches and SVA, surrogate variable analysis.

Upon establishing the negligible batch effect in adult mosquito samples, PCA was performed on the dataset (Figure 3.4-7-A) to observe major variances in the data and if sex was represented by these variances. A total of 48 components were required to explain 95% of variance in the data. The first two components, PC1 (30.11%) and PC2 (14.02%), explained a cumulative variance of 44.13%. This does not distinguish sex as a major source of variance in these samples as this was also observed in pupae data. Therefore, a discriminant PLS-DA model was used to extract this information further.

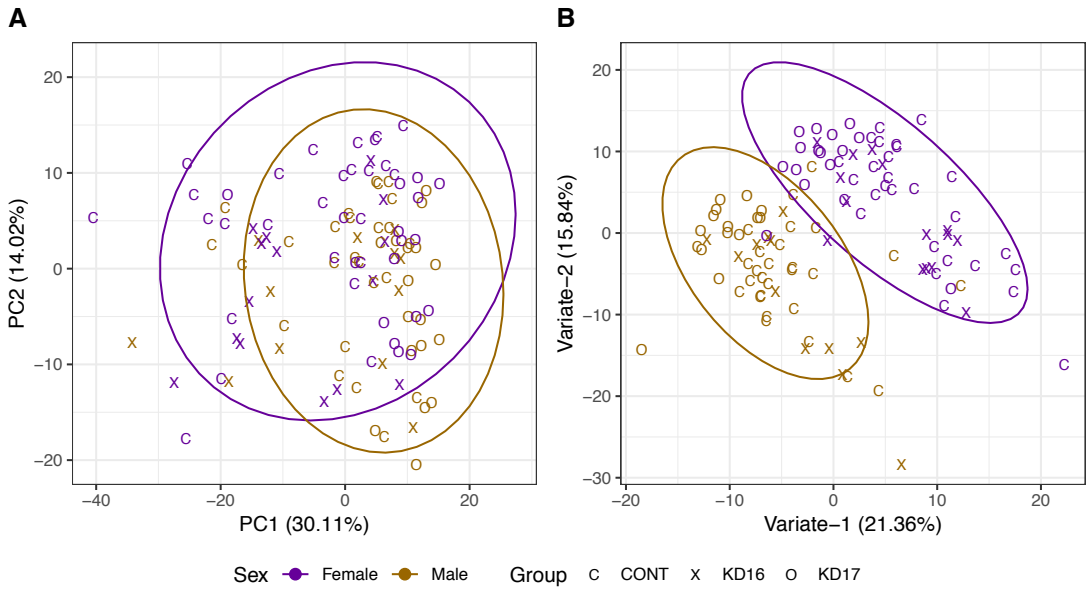


Figure 3.4-7: A) PCA scores of all knock-down *An. gambiae* adult coloured by sex (n_{male} = 54 and n_{female} = 58). Brackets report the variance explained by the PC. Forty-eight PCs were required to achieve 95% explained variance. Only PC1 and PC2 are show in the Figure for simplicity. Sex separation was not observed in the density plots of the remaining PCs. Ellipses represent 95% confidence region. B) PLS-DA scores of adult knock-down of *An.* gambiae discriminated by sex (n_{male} = 54 and n_{female} = 58). Model complexity of two variates was optimised using a cross-validation with 94.12% accuracy. Percentage reported in the brackets report the variance explained by the variate of the model. Ellipses represent 95% confidence region.

Using a cross validated PLS-DA model, sex differences were accentuated as shown in Figure 3.4-7-B. A two-variate PLS-DA model was found to be the optimal complexity with 94.12% accuracy (Appendix 11 for further metrics). As the Figure shows, males and females differentiate clearly along a diagonal of variate-1 and variate-2. The PLS-DA plot also exhibits some samples clustering with the opposite sex. These samples could be the extreme cases of their own groups and were misclassified by the model. This occurrence demonstrates the limitations of the model. Nevertheless, given the low numbers of misclassifications it does not greatly impact the model's success in discriminating between males and females. In order to extract the metabolites the most influential in this discrimination, VIP scores were calculated.

3.4.2.2 Key metabolites between male and female

In order to generate a metabolite list from the PLS-DA model, first spectral features most influential in the discrimination needed to be identified. This was performed by calculating VIP scores from the PLS-DA model. Using a passing score of 1 on variate-1 and variate-2, bins influential in the discrimination between males and females were identified. A total of 109 bins scored higher than the VIP score threshold. Of the 109 bins, only 20 (18.35%) were identified (Figure 3.4-8). These 20 bins were attributed to 8 unique metabolites. Prior to

accepting these metabolites, CRS were calculated. CRS were used to report on a bin representative strength over its metabolite.

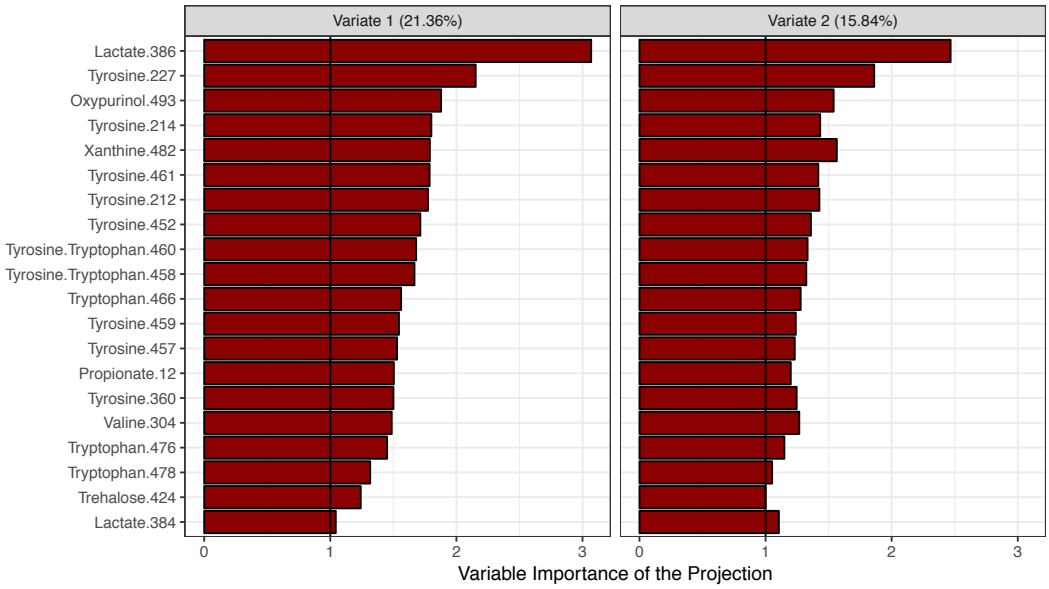


Figure 3.4-8: VIP scores of PLS-DA model built on sex discrimination of adult KD *An. gambiae*. A lower threshold of 1 was used on variates one and two to select metabolites from the model. Black line represents VIP score of 1.

Prior to selecting a metabolite, CRS for all identified bins were calculated in order to ensure the selected bin is a reliable representative of the metabolite. From the CRS calculated from identified bins, a passing score of 29.14% was determined. For the selection process, only bins scoring above the passing score were considered. Only the highest scoring, non-overlapping (where applicable) bins were selected. Table 3.4-3 shows the CRS for the VIP selected bins.

Table 3.4-3: CRS for adult knock-down *An. gambiae*, representing the influential bins from PLS-DA model on sex differences.

Metabolite	Bin	CRS [%]	CRS > 29.14%	Rep	Metabolite	Bin	CRS [%]	CRS > 29.14%	Rep
Lactate	386	49.82	✓	386	Tyrosine	452	70.05	✓	452
	384	35.70	✓			458*	69.82	✓	
Oxypurinol	493	Singlet	NA	493		460*	69.69	✓	
Propionate	12	52.57	✓	12		459	69.45	✓	
Trehalose	424	35.66	✓	424		214	69.21	✓	
Tryptophan	460*	24.00	×	-		457	67.13	✓	
	458*	23.97	×			212	66.66	✓	
	478	20.65	×			461	57.10	✓	
	466	20.09	×			227	22.37	×	
	476	17.66	×			360	-13.88	×	
					Valine	304	-12.94	×	-
					Xanthine	482	Singlet	NA	482

*: Overlapping bin; NA: Singlet self-correlation.

After applying CRS analysis, the list of selected metabolites was shortened to six metabolites (Table 3.4-4). Omitted metabolites were tryptophan and valine. The list of selected metabolites comprised of 4 metabolite classes; amino acids, carboxylic acids, purines and saccharides.

Table 3.4-4: Most influential metabolite list of PLS-DA model for discriminating adult knock-down *An. gambiae* females against males.

Class	Metabolite	Bin	Chemical shift [ppm]	KEGG code
Amino acids	Tyrosine	452	6.90	C00082
Carboxylic acids	Lactate	386	4.13	C00186
	Propionate	12	1.05	C00163
Purines	Oxypurinol	493	8.27	C07599
	Xanthine	482	7.89	C00385
Saccharides	Trehalose	424	5.20	C01083

In order to observe the sex variance represented by the selected metabolites, PCA (Figure 3.4-9-A) was performed on the filtered data to include only the selected metabolites. The PCA scores plot of PC1 (43.81%) against PC3 (15.31%) shows two tightly clustering groups accounting for a cumulative explained variance of 59.12%. A total of 5 components were required to explain 95% variance in the data. The plot exhibits sex-specific variation largely represented by PC1. The intersection of the data clusters shows the limitation of these metabolites in sex discrimination, although, a great portion of the data is clearly separated which can be further accentuated with a supervised PLS-DA model.

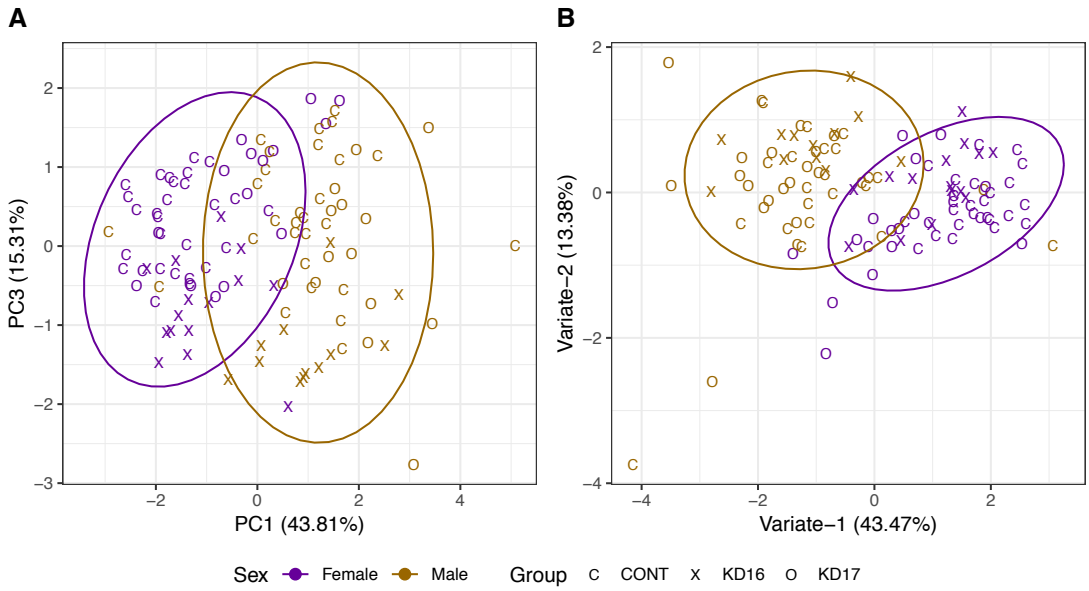


Figure 3.4-9: A) PCA scores for selected metabolites of adult knock-down of *An. gambiae* selected PCA. PC1 (43.81%) and PC3 (15.31%) account for 59.12% cumulative explained variance. A total of 5 components were required to explain 95% variance in the data. Ellipses represent 95% confidence region. B) PLS-DA scores to verify metabolite selection in adult knock-down of *An.* gambiae discriminated by sex (n_{male} = 54 and n_{female} = 58). Model complexity of two-variates explaining 43.47% of the variance was optimised using cross-validation with 94.12% accuracy. Ellipses represent 95% confidence region.

Shortlisted metabolites were used to build a PLS-DA model in order to verify the discriminatory strength of the metabolite used. Cross validated PLS-DA model optimal complexity was determined to be two-variates, with 94.12% accuracy (Appendix 11 for further metrics). The resulting model (Figure 3.4-9-B) shows discrimination to be better than sex differences represented in the previous PCA plot (Figure 3.4-9-A).

Upon metabolite selection of sex-specific differences, metabolite levels were compared *via* BH adjusted t-test (for detailed statistics see Appendix 12) to gain metabolite level information on sex differences. The differences were shown in boxplots (Figure 3.4-10). From the metabolite level comparison, the carboxylic acid propionate was found to be significantly higher in males while lactate was significantly higher in females. The amino acid tyrosine was significantly higher in males as well as the saccharide trehalose. Lastly, the purines xanthine and oxypurinol were found to be significantly higher in males and females respectively.

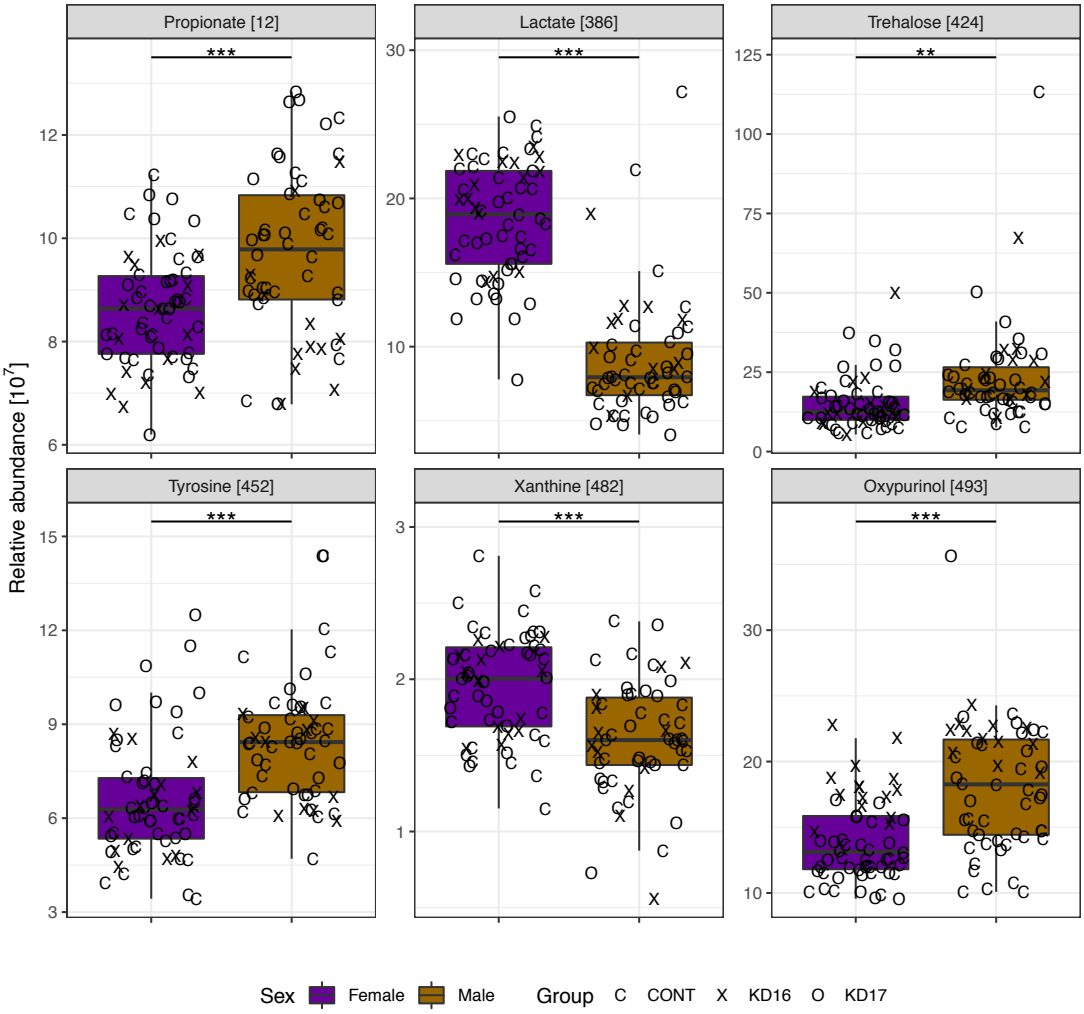


Figure 3.4-10: Boxplots for selected metabolites from adult knock-down *An. gambiae*. ** and *** denotes p-values less than 0.01 and 0.0001 respectively.

3.4.3 Sex differences across stages

Table 3.4-5 shows selected metabolite levels of female species compared to males. Pupa metabolites were not found to be significantly different between sexes. All adult measured metabolites were significantly different in univariate testing, resulting in a metabolic sex profile easier to discriminate in PLS-DA models. In order to understand these changes on a pathway level, MSEA was performed on the elected metabolites.

Table 3.4-5: Metabolite shortlist comprising metabolites selected from the PLS-DA models discriminating females against males for both pupae and adults. Arrows represent significant changes, NS (arrow) denotes non-significant change with mean abundance level, and square brackets report the BH-adjusted p-value.

		Pupa	Adult	
Metabolite class		**Female compared to male**		
Alcohols	Methanol	NS (↑) [3.12×10^{-1}]		
Amino acids	Glycine	NS (↓) [3.12×10^{-1}]		
	Tryptophan	NS (↓) [3.93×10^{-1}]		
	Tyrosine	NS (↑) [8.93×10^{-1}]	↓ [4.14×10^{-6}]	Tyrosine
Carboxylic acids	Acetate	NS (↓) [2.43×10^{-1}]		
	Formate	NS (↓) [3.12×10^{-1}]		
			↑ [1.25×10^{-21}]	Lactate
	Propionate	NS (↓) [3.61×10^{-1}]	↓ [7.98×10^{-5}]	Propionate
Purines	Oxypurinol	NS (↓) [8.93×10^{-1}]	↓ [1.23×10^{-6}]	Oxypurinol
			↑ [3.12×10^{-1}]	Xanthine
Sugars	Glucose	NS (↑) [8.93×10^{-1}]		
			↓ [1.30×10^{-3}]	Trehalose

Selected metabolites were used in MSEA (Figure 3.4-11). Due to the low number of metabolites, only pupae metabolites produced a significantly over-represented pathway (

Table 3.4-6). It should be noted that if the unadjusted raw p-values are considered, the adult metabolite list's only over-represented pathway was propanoate metabolism.

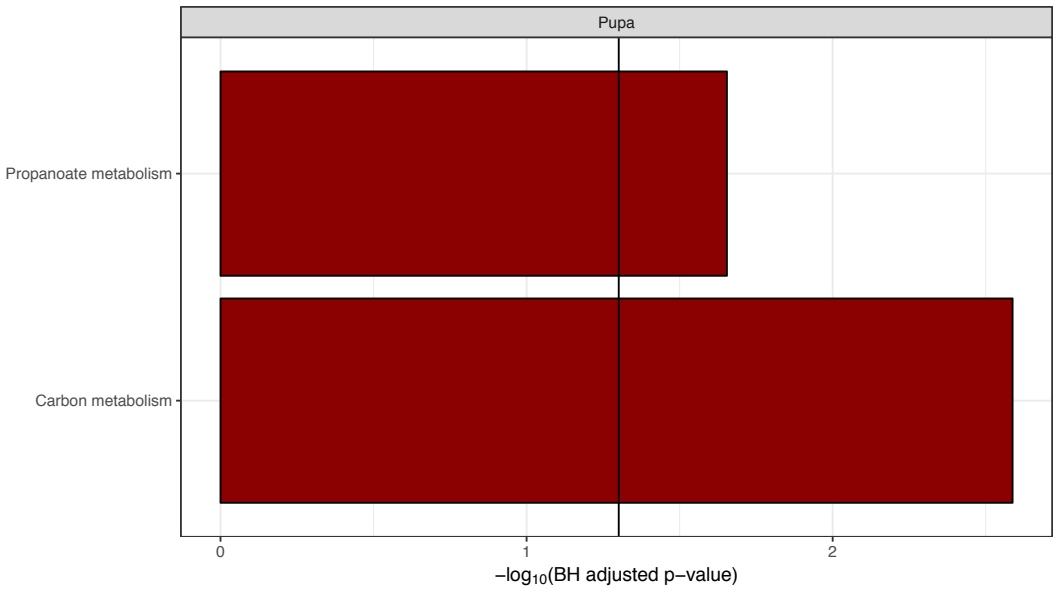


Figure 3.4-11: MSEA of sex differences in knock-down *An. gambiae* pupae. Black line represents p-value of 0.05.

Table 3.4-6: MSEA result details for KD *An. gambiae*, reporting; stage, raw & BH adjusted p-values, number of hits and matched metabolites.

Pathway	Stage	Raw p-value	BH adjusted p-value	Hits/total (%)	Metabolites
Carbon metabolism	Pupa	0.0001	0.0026	4/112 (3.57%)	Acetate, glycine, formate, methanol
Propanoate metabolism	Pupa	0.0014	0.0222	3/48 (6.25%)	Acetate, methanol, propanoate

Following metabolite selection both in pupa and adult stages, a different set of metabolites were shown to be important in differentiating the sexes metabolically. In pupa, nine metabolites were sufficient to explain the differences between sexes in the selected metabolite PCA and PLS-DA model. Whereas, in adults, only five metabolites were required to show differences both in PCA and PLS-DA model. Interestingly, metabolites selected for the pupa model were not found to be significantly different, whereas the lesser numbers of selected metabolites in adults, were all significantly different (lowest being p-value < 0.01). When these were used in MSEA, both pupa (BH-adjusted p-value) and adult (raw p-value) stages over-represented propanoate metabolism.

To summarise, sex-specific differences in knock-down *An. gambiae* species exhibit limited metabolic variance in the data. Although not the major variance, these differences can still be extracted through robust statistical modelling methods. As evidenced by the data, metabolic sex profiles are more similar in pupae compared to in adults. In adults, five metabolites were needed to differentiate between sexes and they were found to be significantly different in the univariate test. Finally, on a pathway level, sex differences are represented by the same pathway both in pupa and adults. This suggests that these sex effects, however minor, can be overcome by employing the same method both in pupae and adults.

3.5 Sex-specific differences in wild type *An. gambiae* pupa and adult metabolic profile

3.5.1 Sex-specific differences in mosquito pupae

3.5.1.1 Statistical analysis

Statistical analysis was started with the assessment of batch effect in the samples. Figure 3.5-1 shows the calculated variance estimation of the data *via* PVCA. For the raw pupa (red bars) dataset, PVCA calculations estimated batch effect to account for 3.55% of the variance.

In comparison to other batch effect correction methods ComBat, (blue bar) performed best in reducing the batch effect variance in the pupa dataset to 0.00%.

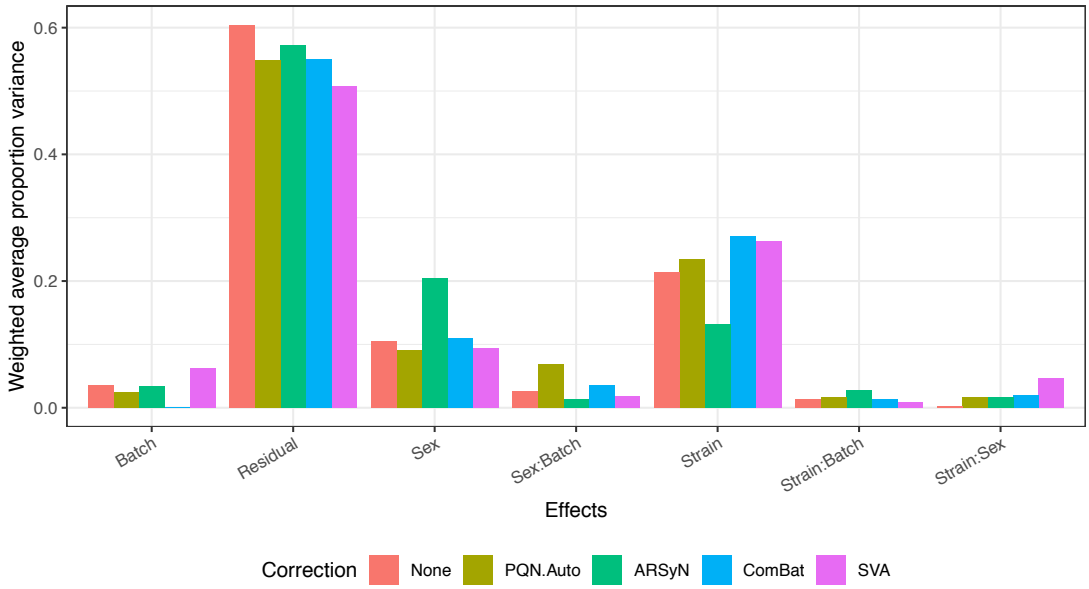


Figure 3.5-1: Estimation of batch, sex and strain variances *via* PVCA in wild type *An. gambiae* pupae. Colon denotes the combined effects of two sources of variance and Residual denotes the total remaining (residual) variation in the data. ComBat outperformed other methods with the lowest batch variance. Correction methods applied: None, no correction; PQN.Auto, probabilistic quotient normalization and autoscaling; ARSyN, ASCA [ANOVA simultaneous component analysis] removal of systematic noise; ComBat, Combining batches and SVA, surrogate variable analysis.

As established in the knock-down analysis, the sex difference for all the comparisons should be considered. To establish possible effects of the knock-down on sexes, wild type mosquitoes were also analysed for sex variation. PCA (Figure 3.5-2-A) was performed to observe the major variances in the data. PCA scores plot of PC1 (22.16%) against PC2 (20.75%) showing weak clustering of sex in terms of separation. PC1 and PC2 explains a cumulative variance of 42.91%. Meanwhile a total of 27 components were required to explain the 95% variance in the data. When the overall metabolic profile of pupae is considered, males are clustered tighter compared to females on PC1 and PC2. This suggests the male population have less variation compared to females. When the largest contributors for the variation is considered, sex is less likely to be one of them. However, the weak clustering observed suggests sex differences can be extracted using a cross-validated model approach.

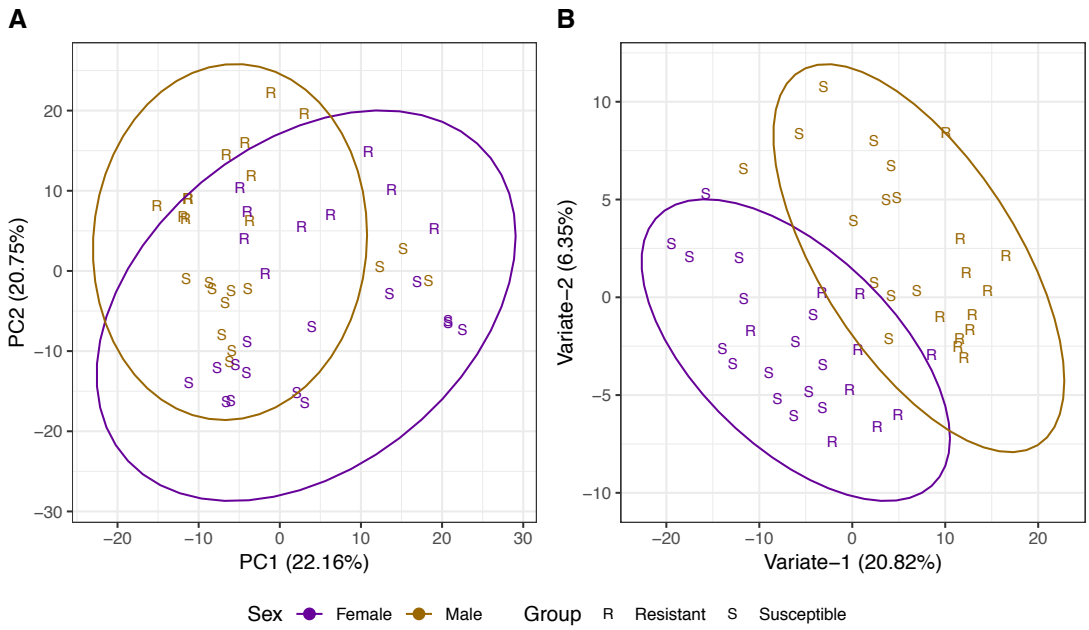


Figure 3.5-2 A) PCA scores plot of wild type *An. gambiae* showing the major variances in the dataset (n_{Female}=24, n_{Male}=33). A total of 27 components were required to explain 95% of the variation in the data. Sex separation was not observed in the density plots of the first 27 components. Ellipses represent 95% confidence region. B) PLS-DA model on sex in wild type *An. gambiae* pupa. The model (n_{Female}=24, n_{Male}=33) was built with four variates and was cross validated with 71.43% accuracy. Brackets represent the explained variance percentage got the given variate. Ellipses represent 95% confidence region.

Although it does not clearly separate males and females, Figure 3.5-2 shows some clustering of the data in terms of sex. This underlying chemical phenotype can be accentuted by the aid of a cross validated PLS-DA model. A cross-validated PLS-DA (Figure 3.5-2-B) model discriminating between females and males was built with four variates achieving 71.43% accuracy (Appendix 11 for further metrics). The PLS-DA scores plot demonstrates the underlying chemical phenotype mentioned in the PCA scores plot. From the scores plot of the PLS-DA model, a separation between males and females can be seen along a diagonal on variate-1 and variate-2. To delve into the metabolite level information, VIP scores of the PLS-DA model were calculated.

3.5.1.2 Key metabolites between males and females

To extract the metabolite level information from the model, VIP scores were calculated. In order to select bins to represent metabolites, only identified bins scoring a VIP score higher than 1 on both variate-1 and variate-2 were selected (Figure 3.5-3). A total of 135 bins scored higher than the threshold. From the 135 bins, only 53 bins (31.26%) were identified and were assigned to 14 unique metabolites. Furthermore, CRS were calculated to assess the reliability of each bin in representing its metabolite.

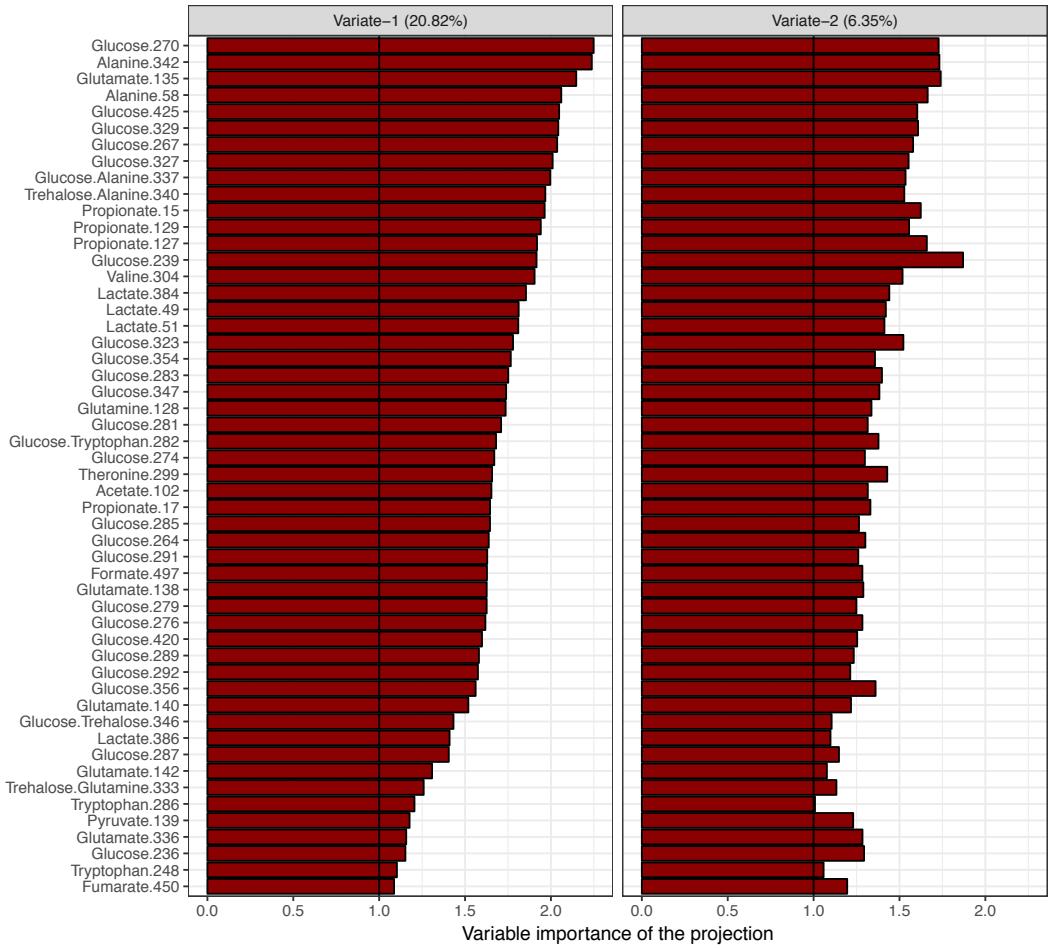


Figure 3.5-3: VIP scores of PLS-DA model built on sex discrimination of wild type *An. gambiae* pupae. A lower threshold of 1 was used on variates one and two to select metabolites from the model. Black line represents VIP score of 1.

Using CRS, bins selected by the VIP criteria were assessed on their reliability on reporting their attributed metabolites. Prior to bin selection, a threshold was calculated using CRS from all identified bins.

Table 3.5-1 shows the CRS calculated for bins selected by the VIP scoring. Only non-overlapping (where applicable) bins scoring higher than the threshold of 31.74% were accepted. Using the CRS table highest scores, non-overlapping (where applicable) bins were selected as the representative of their respective metabolites.

Table 3.5-1: CRS for wild type *An. gambiae* pupae. Only showing CRS for VIP selected bins. Rep: representative bin.

Metabolite	Bin	CRS [%]	CRS > 31.74%	Rep	Metabolite	Bin	CRS [%]	CRS > 31.74%	Rep
Acetate	102	Singlet	NA	102	Glutamate	138	43.40	✓	138
Alanine	337*	81.37	✓	342		140	40.04	✓	
	342	80.89	✓			142	35.01	✓	
	340*	78.94	✓			135	19.68	×	
	58	72.74	✓			336	17.15	×	
Formate	497	Singlet	NA	497	Glutamine	333*	38.12	✓	333
Fumarate	450	Singlet	NA	450		128	27.39	×	
Glucose	270	66.52	✓	270	Lactate	384	90.00	✓	384
	329	65.92	✓			51	87.32	✓	
	281	65.86	✓			49	85.74	✓	
	267	65.76	✓			386	77.89	✓	
	282*	65.72	✓		Propionate	129	64.35	✓	129
	279	65.60	✓			15	60.78	✓	
	283	65.35	✓			17	55.79	✓	
	323	63.40	✓			127	53.79	✓	
	264	62.99	✓		Pyruvate	139	Singlet	NA	139
	425	62.43	✓		Threonine	299	13.82	×	-
	337*	62.39	✓		Trehalose	346*	66.76	✓	346
	327	62.26	✓			333*	56.12	✓	
	285	59.13	✓			340*	20.26	×	
	287	58.87	✓		Tryptophan	248	25.60	×	-
	347	57.66	✓			282*	17.52	×	
	354	56.79	✓			286	-6.70	×	
	420	56.77	✓		Valine	304	34.01	✓	304
	292	56.62	✓						
	356	56.51	✓						
	346*	55.19	✓						
	291	54.31	✓						
	276	53.20	✓						
	289	52.71	✓						
	239	48.43	✓						
	274	46.49	✓						
	236	24.04	×						

From the 14 unique metabolites shortlisted *via* VIP scoring, 12 bins were selected to represent their metabolites (Table 3.5-2). Meanwhile, two metabolites, threonine and tryptophan, were excluded from the final metabolite shortlist due to their low CRS.

Table 3.5-2: Influential metabolites of the PLS-DA model for sex discrimination in wild type *An. gambiae* pupa.

Class	Metabolite	Bin	Chemical shift [ppm]	KEGG code
Amino acids	Alanine	342	3.80	C00041
	Glutamate	138	2.35	C00025
	Glutamine	333*	3.77	C00064
	Valine	304	3.61	C00183
Carboxylic acids	Acetate	102	1.92	C00033
	Formate	497	8.46	C00058
	Fumarate	450	6.52	C00122
	Lactate	384	4.11	C00186
	Propionate	129	2.18	C00163
	Pyruvate	139	2.37	C00022
Sugars	Glucose	270	3.42	C00031
	Trehalose	346*	3.83	C01083

*: Overlapping bin.

Prior to MSEA, the discriminatory properties of these 12 metabolite bins were investigated. Figure 3.5-4-A shows PCA scores plot of PC1 (61.77%) against PC2 (15.82%) for pupae. PC1 and PC2 explain a total of 77.59% of the explained variance, meanwhile, 95% of the variance was explained with six components. In the PCA scores plot, a clear separation of males and females was not observed. Although, along PC1, males and female clusters can be seen suggesting an underlying structure in the data. Applying a cross-validated PLS-DA, the performance of the selected metabolites in discriminating between sexes was assessed.

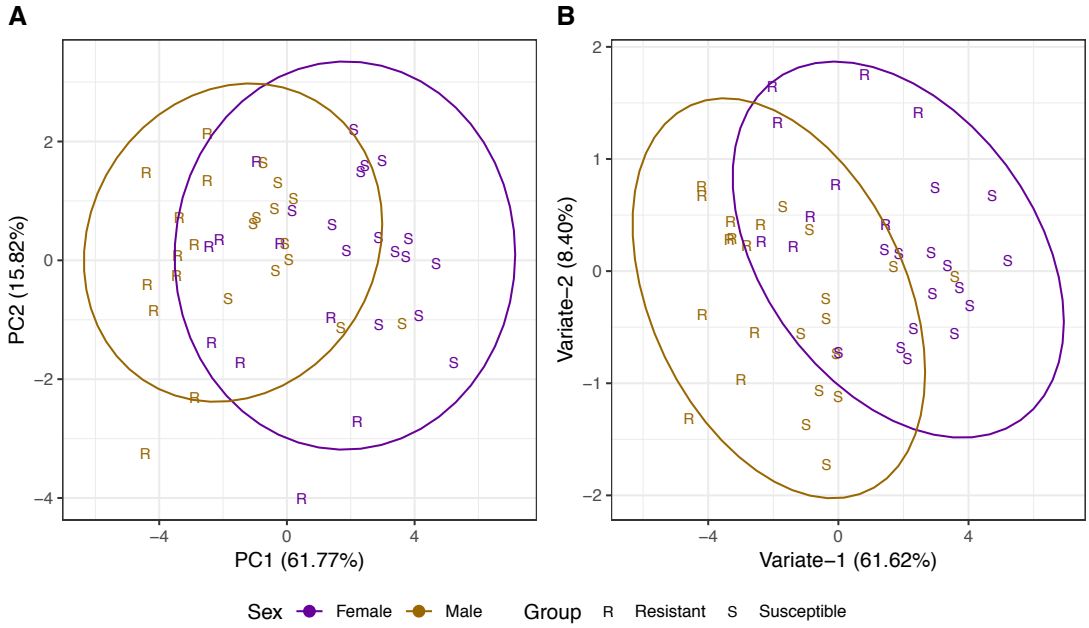


Figure 3.5-4: A) PCA of shortlisted metabolites for wild type *An. gambiae* pupa (n_{Female}=24, n_{Male}=33). PC1 (61.77%) and PC2 (15.82%) account for a cumulative explained variance of (77.59%). A total of six components were required in order to explain 95% of the explained variance. Ellipses represent 95% confidence region. B) PLS-DA scores from selected metabolites discriminating wild type *An. gambiae* pupa by sex (n_{Female}=24, n_{Male}=33). Two-variate model complexity was optimised using cross-validation with 71.43% accuracy. Brackets report the explained variance for the variate. Ellipses represent 95% confidence region.

A PLS-DA model (Figure 3.5-4-B) to discriminate between males and females was built. Through cross-validation, optimal model complexity was determined to be a two-variate model with 71.43% accuracy (Appendix 11 for further metrics). PLS-DA scores plot show a greater degree of separation compared to the PCA scores plot. Discrimination of females against males can be seen along a diagonal of variate-1 and variate-2. Intersecting data points demonstrate the similarities between males and females as well as the limitation of the model in discriminating between males and females. To obtain metabolite level information, BH adjusted t-test (for detailed statistics see Appendix 12) was used to compare the metabolite levels. These comparisons were demonstrated with boxplots (Figure 3.5-5).

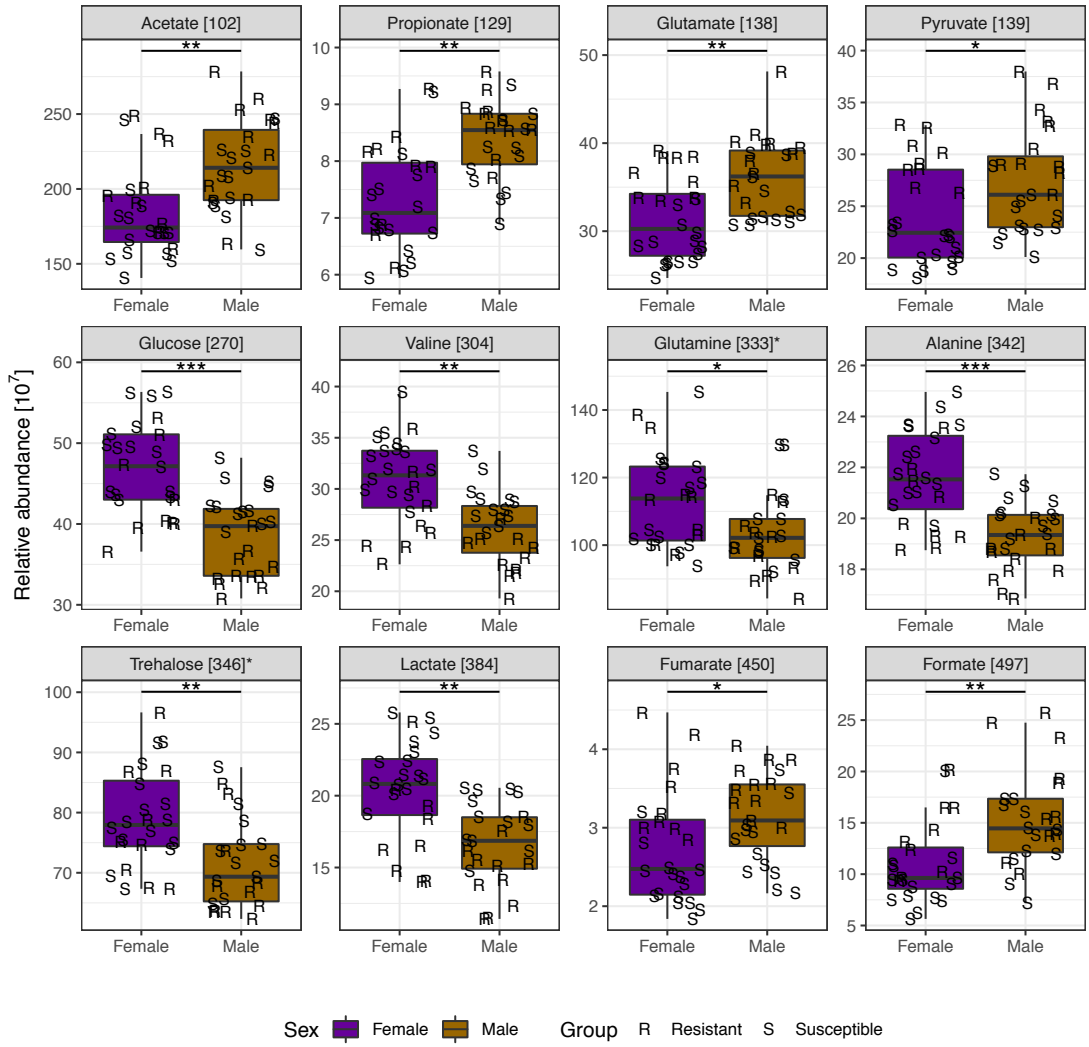


Figure 3.5-5: Selected metabolite boxplots of male (n_{Male}=33) and female (n_{Female}=24) wild type *An. gambiae* pupae. *, ** and *** represent p-value less than 0.05, 0.01 and 0.001 respectively. * in the boxplot title represent denotes overlapping bin.

Univariate analysis revealed selected metabolites representing the sex differences to be significantly different between sexes. Within the selected metabolites, representative bins for glutamine and trehalose were of overlapping peaks. In the comparison of selected metabolites, all the amino acids (valine, glutamine and alanine) were found to be significantly higher in females except glutamate where it was higher in males. All carboxylic acids (acetate, propionate, pyruvate, fumarate and formate) were significantly higher in males except lactate where it was higher in females. Lastly, all saccharides (glucose and trehalose) were higher in females.

3.5.2 Sex-specific differences in adult mosquitoes

3.5.2.1 Statistical analysis

Prior to the statistical analysis of sex differences, variation types in the data were estimated using PVCA. Figure 3.5-6 shows the variation estimation in the data. As the figure shows, there is no batch effect seen in the adult dataset. Therefore, the data was not treated with a batch correction method and was PQN normalised followed by auto scaling prior to statistical analysis. It should be noted that the PQN normalisation and auto scaling increases the batch effect contribution to 0.45% but, compared to the 18.64% sex effect, it is negligible.

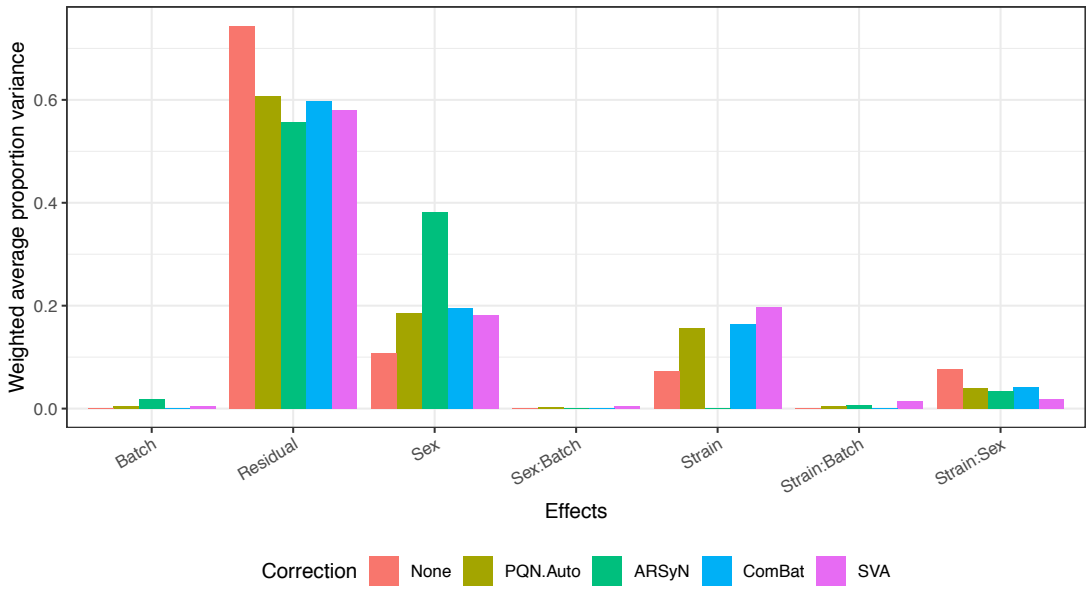


Figure 3.5-6: Estimation of batch, sex and strain variances *via* PVCA in wild type *An. gambiae* adult. Colon denotes the combined effects of two sources of variance and Residual denotes the total remaining variation in the dataset (residua). Data shows no batch effect. Correction methods applied: None, no correction; PQN.Auto, probabilistic quotient normalization and autoscaling; ARSyN, ASCA [ANOVA simultaneous component analysis] removal of systematic noise; ComBat, Combining batches and SVA, surrogate variable analysis.

Following the variation estimate, PCA was performed to observe the major variances in the data. The PCA scores plot (Figure 3.5-7-A) does not show a clear separation between males

and females. Although a clear separation was not observed along PC1 (38.67%), male samples are clustered tighter compared to females, suggesting a higher degree of variability within female species. In contrast, sex difference was observed with better resolution along PC3 (8.63%). PC1 and PC3 accounted for a total variance of 47.30% and 29 components were required to explain the 95% cumulative variance. Along PC1, a female sub-population can be observed, although, given the metadata this could not be attributed to any known factor of batch and sex. Upon the observation of the loadings of PC1 (Appendix 13), glucose demonstrates a high influence in the separation observed. This is most likely due to the feeding habits of the mosquitoes. This underlying structure for sex differences, especially exhibited on PC3, can be further used in a supervised PLS-DA for further enhancement of the sex differences and identification of the metabolite differences observed.

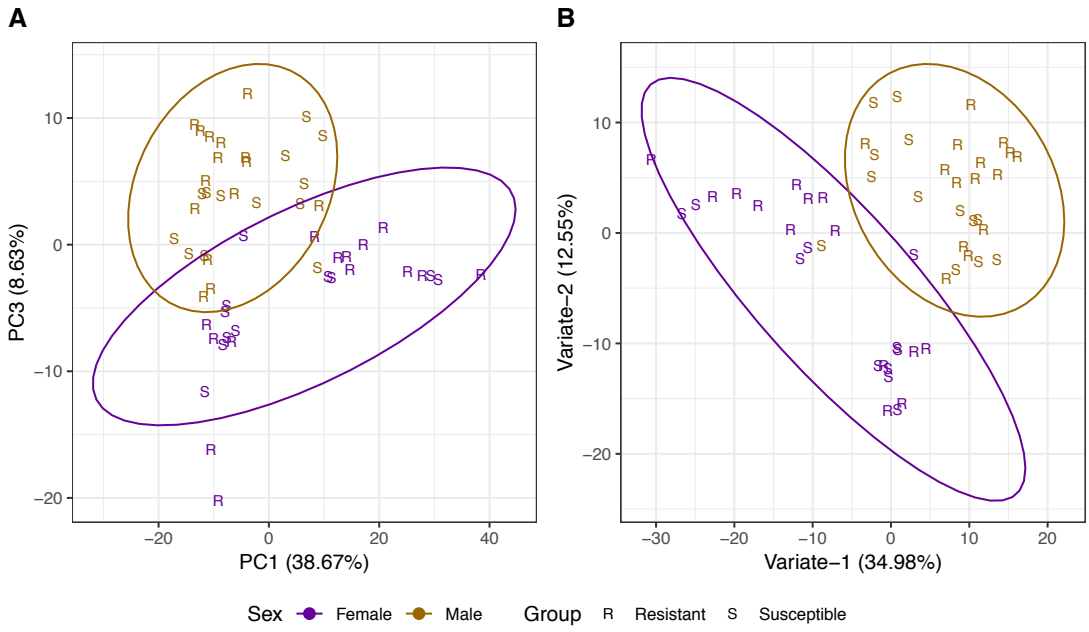


Figure 3.5-7: A) PCA scores plot of adult wild type *An. gambiae* coloured by sex (n_{Female}=25, n_{Male}=28). A total of 29 components were required for 95% explained variance. PC1 (38.91%) shows a higher degree of variation within females. Meanwhile, the most pronounced sex separation was observed on PC3 (8.56%). Ellipses represent 95% confidence region. B) PLS-DA scores of adult wild type *An. gambiae* discriminated by sex (n_{Female}=25, n_{Male}=28). Model complexity of two variates was optimised using cross-validation with 87.50% accuracy. Ellipses represent 95% confidence region.

A cross-validated PLS-DA model (Figure 3.5-7-B) was built, which was then used for metabolite selection. Through the cross-validation, a two-variate model with 87.50% accuracy (Appendix 11 for further metrics) was found to be the optimal complexity for sex discrimination for adult wild type *An. gambiae*. PLS-DA scores plot retained the sub-populations observed in the PCA scores plot (Figure 3.5-7-A), which could not be attributed to any measured factor. Scattering of the sub-population suggests variate-1 and variate-2

were used to discriminate between the two female sub-populations and males. Using VIP scores, this discrimination can be probed further for metabolite level information.

3.5.2.2 Key metabolites of the comparison

VIP scores were calculated for all the bins in the model. In order to select metabolites influential in discriminating between males and females, identified bins scoring higher than 1 were selected. A total of 179 bins out of 496 (36.09%) scored above the VIP score threshold of one. From the 179 bins, only 58 (32.40%) were identified and were attributed to 14 unique metabolites (Figure 3.5-8). In order to select metabolites, these selections from VIP scoring needed to be further checked for their representative quality for their respective metabolites, which was performed by CRS.

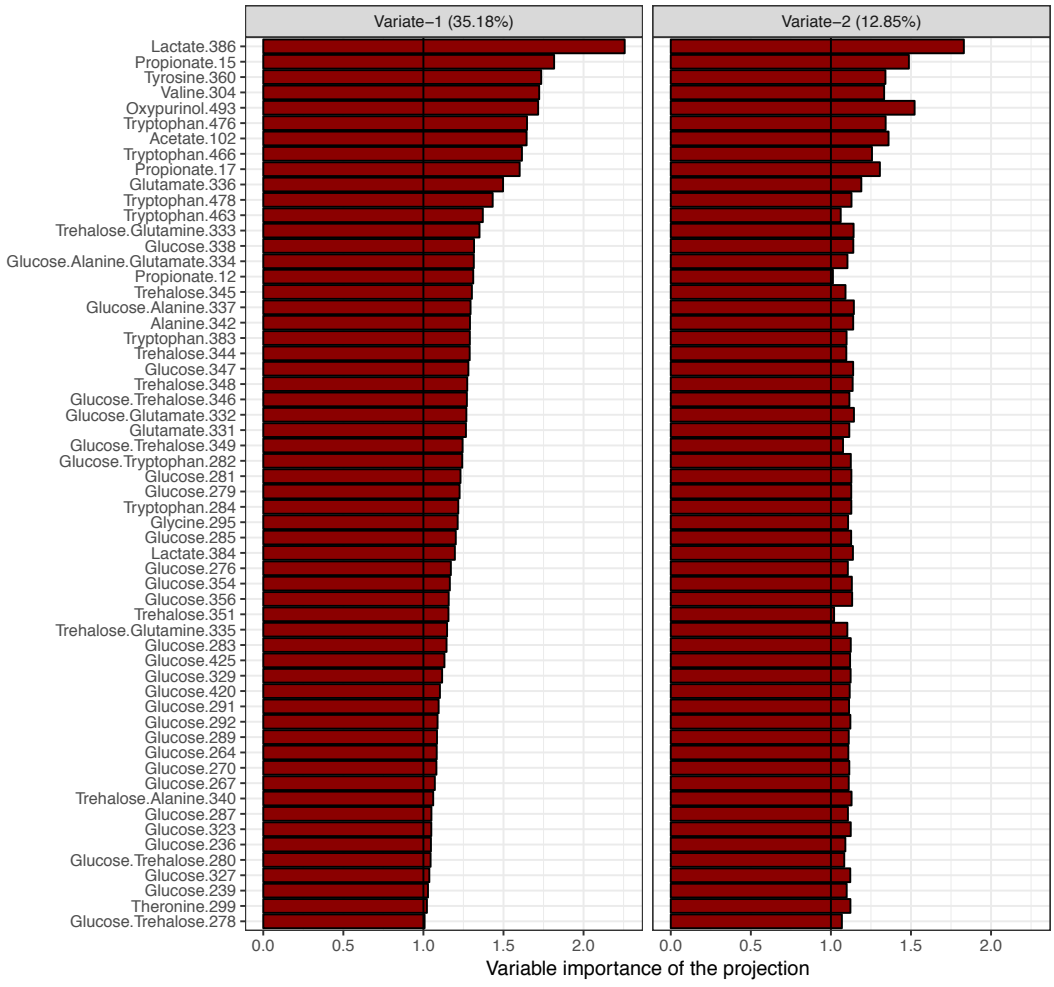


Figure 3.5-8: VIP scores of wild type *An. gambiae* adult PLS-DA model discriminating between males and females. A total of 58 bins out of 496 (11.69%) were selected representing 14 unique metabolites. Black line represents VIP score of 1.

CRS was calculated for all identified metabolites. A passing threshold of 35.82% was calculated using all the CRS. Each score for the VIP selected bins were compared to the threshold (Table 3.5-3). Passing bins were taken into consideration for representing their respective metabolites. Out of the passing scores, the highest non-overlapping (where applicable) scores were selected as the representative bin of the metabolite.

Table 3.5-3: CRS for wild type adult *An. gambiae*. A passing score of 35.82% is required for a bin to be considered as a representative of a metabolite.

Metabolite	Bin	CRS [%]	CRS > 35.82%	Rep	Metabolite	Bin	CRS [%]	CRS > 35.82%	Rep
Acetate	102	Singlet	NA	102	Glutamate	334*	13.97	×	-
Alanine	337*	79.78	✓	342		331	12.37	×	
	342	79.58	✓			332*	10.95	×	
	334*	77.28	✓			336	6.80	×	
	340*	75.69	✓		Glutamine	333*	22.20	×	-
Glucose	354	98.00	✓	354		335*	21.83	×	
	356	97.97	✓		Glycine	295	Singlet	NA	295
	283	97.78	✓		Lactate	386	46.84	✓	386
	292	97.71	✓			384	30.28	×	
	285	97.61	✓		Oxypurinol	493	Singlet	NA	493
	329	97.47	✓		Propionate	12	58.17	✓	12
	291	97.38	✓			15	54.93	✓	
	425	97.35	✓			17	54.58	✓	
	270	97.26	✓		Threonine	299	-0.23	×	-
	420	97.24	✓		Trehalose	351	90.94	✓	351
	279	97.24	✓			335*	90.73	✓	
	289	97.22	✓			278*	90.53	✓	
	287	97.19	✓			348	90.24	✓	
	264	97.16	✓			280*	90.19	✓	
	267	97.15	✓			333*	88.77	✓	
	239	97.10	✓			349*	88.61	✓	
	327	97.08	✓			346*	86.56	✓	
	323	97.01	✓			340*	85.34	✓	
	281	96.92	✓			344	84.14	✓	
	236	96.89	✓			345	82.82	✓	
	332*	96.69	✓		Tryptophan	476	21.85	×	-
	280*	96.65	✓			463	21.12	×	
	347	96.64	✓			478	20.41	×	
	337*	96.58	✓			466	19.98	×	
	282*	96.47	✓			284	-2.07	×	
	338	96.02	✓			282*	-2.26	×	
	278*	95.72	✓			383	-2.86	×	
	276	95.35	✓		Tyrosine	360	-7.91	×	-
	346*	94.43	✓		Valine	304	-14.50	×	-
	349*	91.76	✓						
	334*	90.44	✓						

* Overlapping bin

VIP scoring shortlisted 14 unique metabolites. When CRS was applied, this list was further shortened to 8 metabolites. Excluded metabolites included threonine, tryptophan, tyrosine

and valine. Table 3.5-4 shows the metabolites shortlisted from VIP scoring and CRS. The metabolite shortlist comprises four metabolite classes: amino acids (alanine and glycine), carboxylic acids (acetate, lactate and propionate), purines (oxypurinol) and saccharides (glucose and trehalose).

Table 3.5-4: Selected metabolites of PLS-DA model discriminating between males and females in wild type adult *An. gambiae*.

Class	Metabolite	Bin	Chemical shift [ppm]	KEGG code
Amino acids	Alanine	342	3.80	C00041
	Glycine	299	3.57	C00037
Carboxylic acids	Acetate	102	1.92	C00033
	Lactate	386	4.13	C00186
	Propionate	12	1.05	C00163
Purines	Oxypurinol	493	8.27	C07599
Saccharides	Glucose	354	3.89	C00031
	Trehalose	351	3.87	C01083

Prior to performing MSEA, discriminatory properties of only the selected metabolites were assessed. This was performed by filtering the data to include only the representative bins of the selected metabolites, then applying PCA. Using the filtered data, PCA scores plots (Figure 3.5-9-A) show separation between males and females along PC2 (16.02%) which is augmented by PC1 (67.06%). The same female sub-population can still be observed along PC1 and the large variance explained by this component suggest it is caused by the observed female sub-population. PC1 and PC2 accounted for a cumulative variance of 83.08%, meanwhile four components were required to explain the 95% variance in the filtered data. A supervised PLS-DA model was also built in order to determine the performance of these metabolites in discriminatory models.

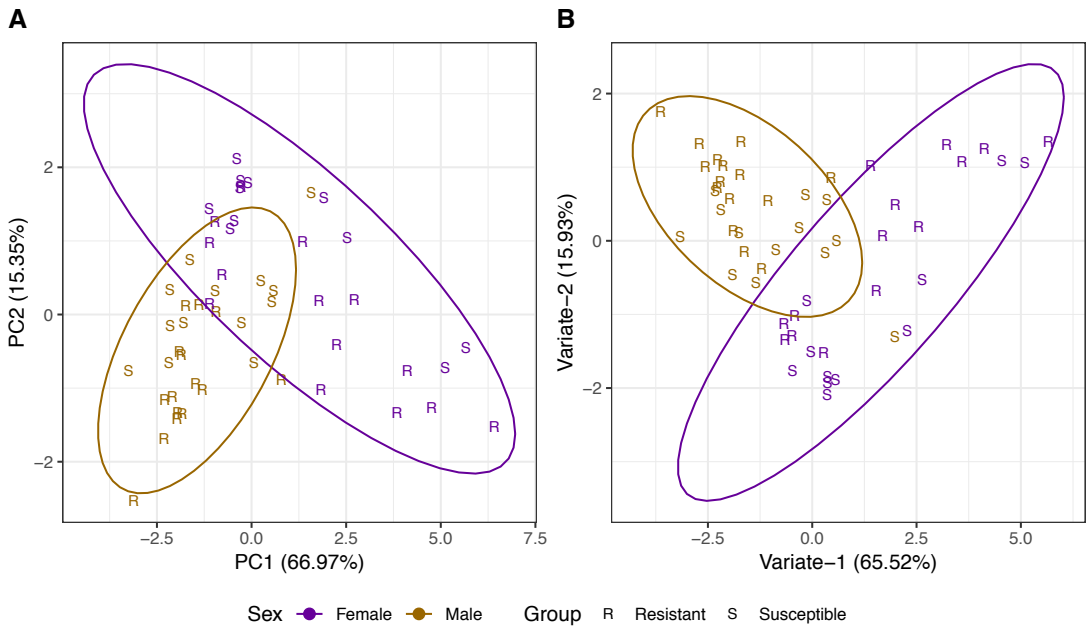


Figure 3.5-9: A) PCA scores plot PC1 against PC2 of wild type adult *An. gambiae* from selected metabolites (n_{Female}=25, n_{Male}=28). PC1 and PC2 accounts for a cumulative explained variance of 83.08%. Meanwhile, four components were required to explain the 95% variance. Ellipses represent 95% confidence region. B) PLS-DA scores of adult wild type *An. gambiae* discriminated by sex (n_{Female}=25, n_{Male}=28). Model complexity of three variates was optimised using cross-validation with 93.75% accuracy. Brackets report the explained variance for the variate. Ellipses represent 95% confidence region.

Selected metabolites were also used to build a PLS-DA model in order to assess their performance on discrimination models exclusively. A cross-validated model was built where a two-variate model was determined to be optimal with 93.75% accuracy (Appendix 11 for further metrics). PLS-DA scores plot (Figure 3.5-9-B) shows a separation between sexes along a diagonal of variate-1 and variate-2. Although there are some classification errors which are easily recognised by data point clustering with the opposite sex, the majority could be classed correctly. Metabolite comparison was carried out in order to gain metabolite level insight on the divergent pathways between males and females.

Metabolite levels were shown *via* boxplots (Figure 3.5-10) and significance were calculated using BH adjusted t-test (for detailed statistics see Appendix 12). All selected metabolites were found to be significantly different between males and females. More specifically, selected amino acids and sugars were all found to be significantly higher in females. The only purine selected was oxypurinol and it was found to be significantly higher in males. In carboxylic acids, while acetate and propionate were found to be significantly higher in males, lactate was significantly higher in females.

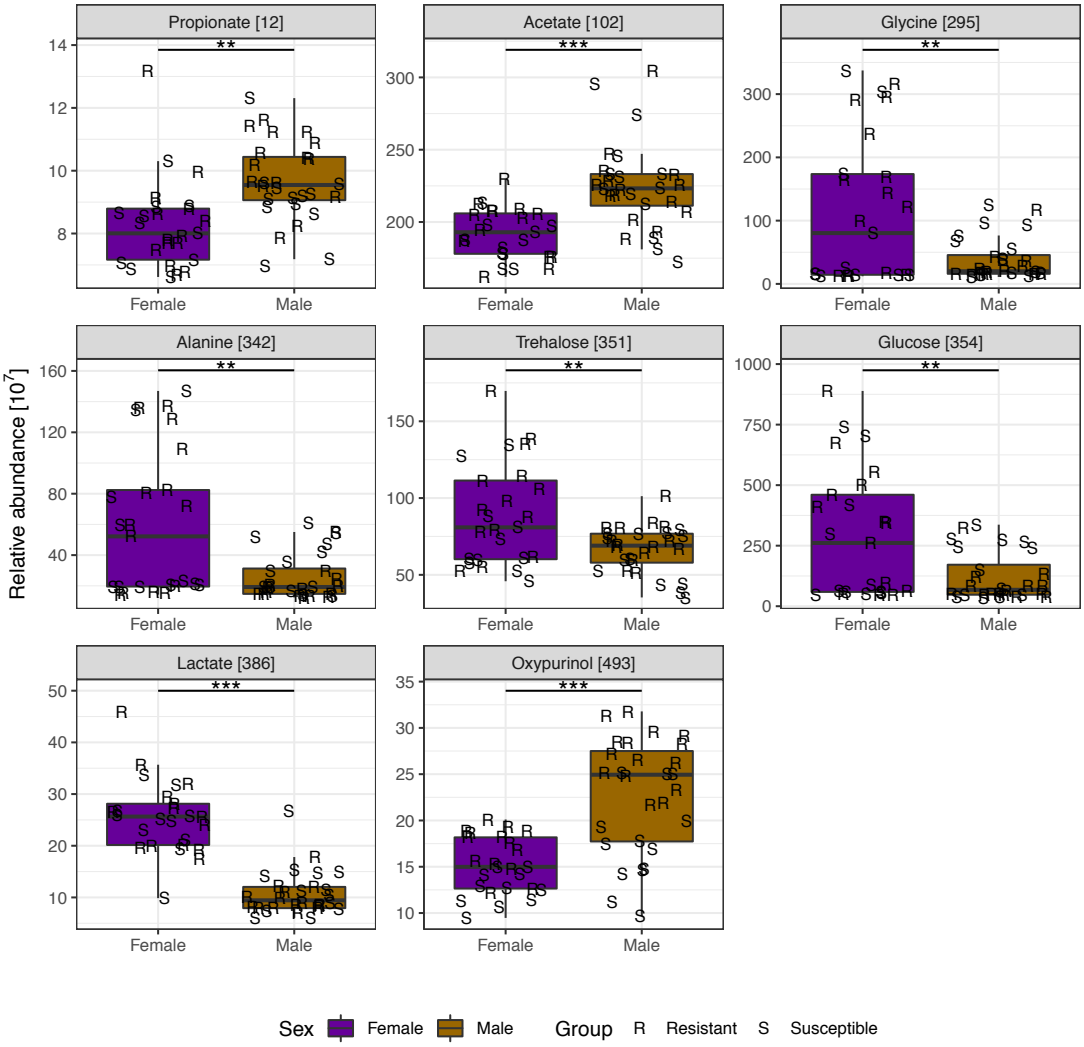


Figure 3.5-10: Boxplots of selected metabolites from PLS-DA model discriminating between females and males in adult wild type *An. gambiae* (n_{Female}=25, n_{Male}=28). ** and *** represents p-values less than 0.01 and 0.001 respectively.

3.5.3 Sex-specific differences across stages

Metabolite level differences were summarised in Table 3.5-5. The table clearly shows metabolite selection yielded 12 metabolites in pupa compared to the 8 metabolites in adults. It should be noted that overlapping metabolites both in pupa and adults follow the same trend across stages. PLS-DA scores plots have shown the metabolic profiles regarding sexes could be discriminated more successfully in adults. Interestingly, adults required a smaller number of unique metabolites while achieving better discrimination between sexes.

Table 3.5-5: Metabolite changes summary for wild type *An. gambiae* pupae and adults. Arrows represent significant metabolite level differences when females were compared to males, and square brackets report BH-adjusted p-values.

	Pupa		Adult	
Class	**Female compared to male**			
Amino acids	Alanine	↑ [4.00×10^{-5}]	↑ [2.74×10^{-3}]	Alanine
	Glutamate	↓ [1.82×10^{-3}]		
	Glutamine	↑ [1.09×10^{-2}]		
			↑ [4.38×10^{-3}]	Glycine
	Valine	↑ [3.43×10^{-4}]		
Carboxylic acids	Acetate	↓ [1.82×10^{-3}]	↓ [3.04×10^{-5}]	Acetate
	Formate	↓ [2.09×10^{-3}]		
	Fumarate	↓ [3.57×10^{-2}]		
	Lactate	↑ [3.43×10^{-4}]	↑ [3.63×10^{-10}]	Lactate
	Propionate	↓ [2.53×10^{-4}]	↓ [1.08×10^{-3}]	Propionate
	Pyruvate	↓ [2.17×10^{-2}]		
Purines			↓ [3.46×10^{-6}]	Oxypurinol
Sugars	Glucose	↑ [4.00×10^{-5}]	↑ [4.38×10^{-3}]	Glucose
	Trehalose	↑ [2.82×10^{-3}]	↑ [4.38×10^{-3}]	Trehalose

Using the selected metabolites for both pupae and adults, a MSEA was performed (Figure 3.5-11). High numbers of selected metabolites from the pupae resulted in a longer list of over-represented pathways (Table 3.5-6). In contrast, adult metabolite selection produces a much more concise list of pathways over-represented. Considering the more successful discrimination achieved in adult mosquitoes, glycolysis/gluconeogenesis and propanoate metabolism may be a better representative for sex separation in both pupae and adults. It should be noted that if the unadjusted raw p-values are considered, propanoate metabolism in pupa was also significantly over-represented.

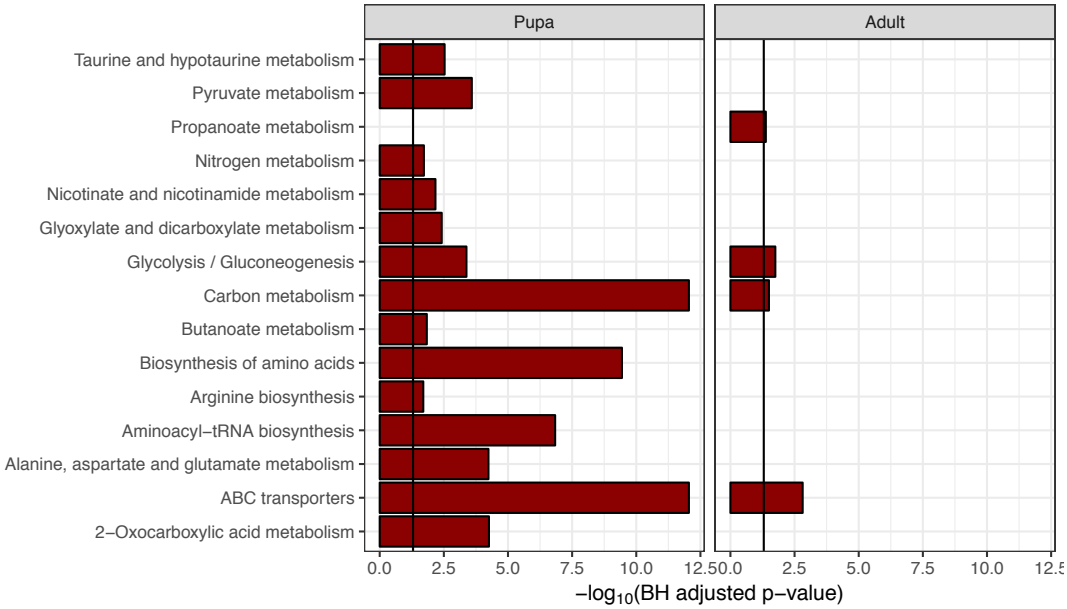


Figure 3.5-11: MSEA from selected metabolites in wild type *An. gambiae* pupae and adults. Black line represents p-value of 0.05.

Table 3.5-6: MSEA result details for wild type *An. gambiae*, reporting; stage, raw & BH adjusted p-values, number of hits and matched metabolites.

Pathway	Stage	Raw p-value	BH adjusted p-value	Hits/total (%)	Metabolites
2-Oxocarboxylic acid metabolism	Pupa	6.0×10^{-6}	5.5×10^{-5}	3/134 (2.24%)	Pyruvate, glutamate, valine
ABC transporters	Pupa	3.9×10^{-14}	8.9×10^{-13}	6/182 (3.30%)	Glutamate, glucose, alanine, glutamine, valine, trehalose
	Adult	5.0×10^{-5}	1.5×10^{-3}	4/182 (2.20%)	Glucose, glycine, alanine, trehalose
Alanine, aspartate and glutamate metabolism	Pupa	7.5×10^{-6}	5.8×10^{-6}	5/29 (17.24%)	Pyruvate, glutamate, alanine, glutamine, fumarate
Aminoacyl-tRNA biosynthesis	Pupa	1.3×10^{-8}	1.5×10^{-7}	4/52 (7.69%)	Glutamate, alanine, glutamine, valine
Arginine biosynthesis	Pupa	6.1×10^{-3}	2.0×10^{-2}	3/23 (13.04%)	Glutamate, glutamine, fumarate
Biosynthesis of amino acids	Pupa	2.4×10^{-11}	3.6×10^{-10}	5/128 (3.91%)	Pyruvate, glutamate, alanine, glutamine, valine
Butanoate metabolism	Pupa	3.8×10^{-3}	1.5×10^{-2}	3/42 (7.14%)	Pyruvate, glutamate, fumarate
Carbon metabolism	Pupa	3.9×10^{-14}	8.9×10^{-13}	6/112 (5.36%)	Pyruvate, glutamate, acetate, alanine, formate, fumarate
	Adult	3.2×10^{-3}	3.2×10^{-2}	3/112 (2.68%)	Acetate, glycine, alanine
Glycolysis / Gluconeogenesis	Pupa	7.0×10^{-5}	4.5×10^{-4}	4/31 (12.90%)	Pyruvate, glucose, acetate, lactate
	Adult	1.2×10^{-3}	1.8×10^{-2}	3/31 (9.68%)	Glucose, acetate, lactate
Glyoxylate and dicarboxylate metabolism	Pupa	8.3×10^{-4}	3.8×10^{-3}	4/64 (6.25%)	Pyruvate, glutamate, formate, glutamine
Nicotinate and nicotinamide metabolism	Pupa	1.6×10^{-3}	6.7×10^{-3}	3/55 (5.45%)	Pyruvate, fumarate, propanoate
Nitrogen metabolism	Pupa	5.3×10^{-3}	1.9×10^{-2}	3/43 (6.98%)	Glutamate, formate, glutamine
Propanoate metabolism	Adult	5.6×10^{-3}	4.1×10^{-2}	3/48 (6.25%)	Acetate, propanoate, lactate
Pyruvate metabolism	Pupa	4.0×10^{-5}	2.6×10^{-4}	5/31 (16.13%)	Pyruvate, acetate, formate, fumarate, lactate
Taurine and hypotaurine metabolism	Pupa	5.8×10^{-4}	3.0×10^{-3}	4/22 (18.18%)	Pyruvate, glutamate, acetate, alanine

The analyses showed the sex differences in pupae although present, are not a major variation in the dataset. By means of statistical modelling and variable selection, these differences can be extracted and further used for different purposes. These differences are more pronounced in adults and some precaution should be taken. When sex differences were being probed, pupal metabolic profiles of males and females were found to be more similar compared to the adult profiles. Both in PCA and PLS-DA approaches, sex differences in pupae were less pronounced and harder to discriminate. In PCA scores plots, neither pupae or adults achieved a clear separation of sexes but nevertheless, adult scores plots showed tighter sex clustering. Additionally, in adults, all dataset PLS-DA models required two variates to achieve 87.50% accuracy whereas pupae PLS-DA model scored 71.43% accuracy with two variates, demonstrating the challenges in discriminating chemical phenotypes of pupae

compared to adults. When a metabolite selection approach was employed, adults required less metabolites to achieve good discrimination compared to pupae. During the univariate comparison, common metabolites selected in pupae and adults followed the same abundance levels across stages (Table 3.5-5), suggesting sex differences can be further refined to a specific set of common metabolites.

3.6 Sex-specific differences in wild type *Ae. aegypti* pupae and adult metabolic profile

3.6.1 Sex-specific differences in mosquito pupae

3.6.1.1 Statistical analysis

Prior to assessing sex differences batch variation in pupal data was assessed using a PVCA estimation (Figure 3.6-1). Variation estimations show batch effects account for 0.64% of the variance in the data, contributing more than the threshold of 0.5%. However, the batch effects could not be fully removed by normalisation and scaling. Therefore, the batch correction method of ComBat was selected reducing the batch contribution to 1.47 x 10^{-19}%, ensuring the variation in the data was not affected by batch variation. Following this assessment, the PCA was performed in order to observe the major variances in the data.

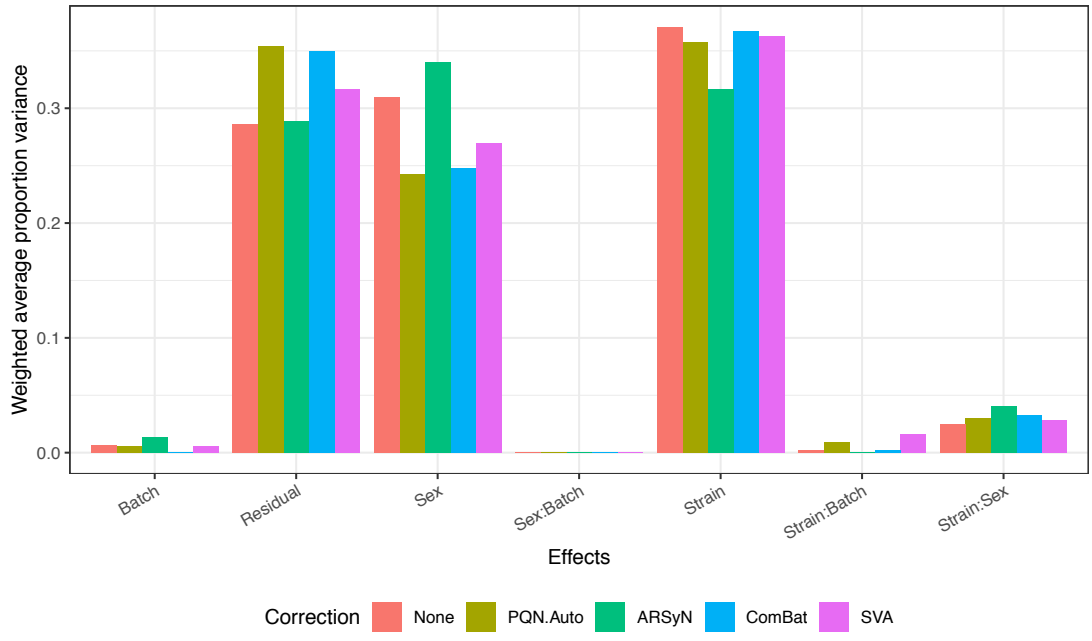


Figure 3.6-1: PVCA of wild type *Ae. aegypti pupa* (n_{Female}=26, n_{Male}=29). Each bar colour represents a different data treatment method. Each batch correction method was normalised before correction and scaled after with PQN normalisation and auto scaling respectively. Where effects of two factors were estimated ':' was used to show it. Residual, remaining total variation (residual) in data; PQN.Auto, PQN normalisation and Auto scaling.

In order to observe the major variance in the data, PCA was performed. The PCA scores plot (Figure 3.6-2-A) of PC1 (33.39%) against PC2 (14.82%) accounted for 48.21% variance in the data. A total of 29 components were required in order to explain the 95% variance in the data. The results shows a clear separation between males and females, except one male sample. This is most likely an extreme case sample, since this sample was not classed as an outlier given the sample passed all QC criteria. Within the plot, two sub groups can be seen in male and female which perfectly match with the resistance status of the mosquitoes. To minimise these differences and accentuate the difference between males and females, a cross-validated PLS-DA model was used. This allowed metabolite selection representing the differences between the two sexes.

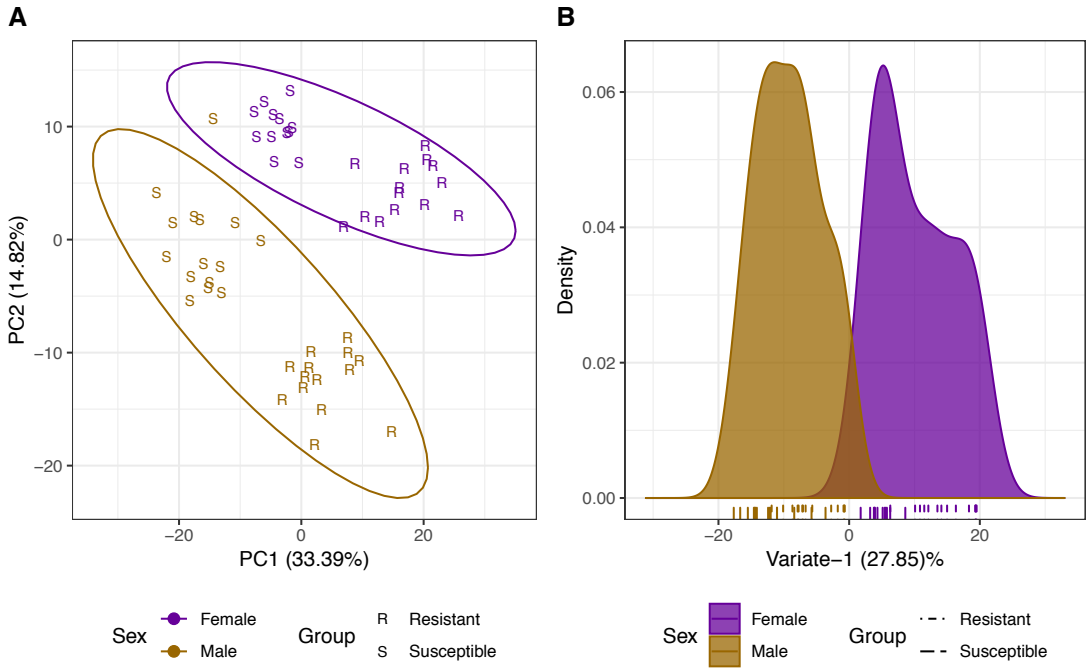


Figure 3.6-2: A) PCA scores plot of wild type *Ae. aegypti* pupa coloured by sexes (n_{Female}=26, n_{Male}=29). PC1 (33.39%) against PC2 (14.82%) accounting for a cumulative varinace of 48.21%. A total of 29 componenets were required in order to explain 95% of variance in the data. Ellipses represent the 95% confidence region. B) PLS-DA density plot showing discrimination of females against males (n_{Female}=26, n_{Male}=29). Each tick represents a sample coloured by its group. Classification accuracy of the PLS-DA model is 93.75%. Brackets report the explained variance for the variate.

A PLS-DA model with cross-validation was built in order to discriminate between male and female mosquitoes (Figure 3.6-2-B). Cross-validation results show the optimal complexity to be a single-variate model with 93.75% accuracy (Appendix 11 for further metrics). In order to understand the differences on a metabolite level, VIP scores were calculated.

3.6.1.2 Key metabolites of the comparison

VIP scores were calculated to select the bins that are most influential on the discrimination between males and females. VIP scores were calculated for all the bins and only identified bins that scored higher than the threshold of 1 were selected. A total of 209 out of 513 bins scored above the VIP threshold of 1. Out of the 209, only 60 (28.71%) were identified and attributed to 15 metabolites (Figure 3.6-3). In order to generate a list of metabolites, the reliability of the bins in terms of representing its metabolite was evaluated by calculating CRS.

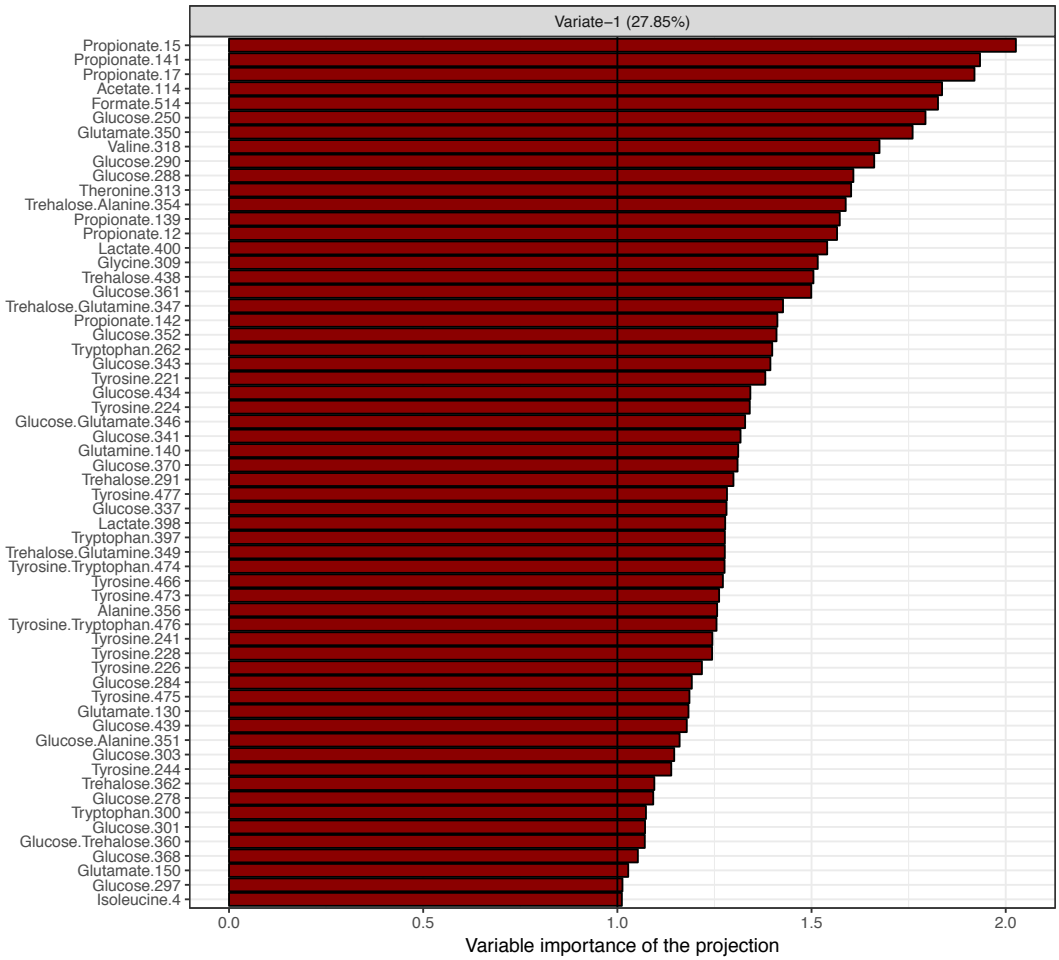


Figure 3.6-3: VIP scores from PLS-DA model of wild type *Ae. aeypti* pupae discirminating between males and females. Only 60 identified bins scored higher than the threshold of 1 out of 513 bins. Selected bins represent a total of 15 metabolites. Black line represents the VIP score of 1.

CRS (Table 3.6-1) was calculated for all the bins in order to establish the threshold score. Threshold score of 42.82% was calculated from all the scores. This threshold was applied to the bins selected from VIP scores. Only non-overlapping (where applicable) bins scoring higher than this threshold were considered as a possible representative of the metabolite. Only the highest scoring bins were selected amongst the bins meeting the scoring criteria.

Table 3.6-1: CRS for *Ae. aegypti* pupa. Rep: representative bin.

Metabolite	Bin	CRS [%]	CRS > 42.82%	Rep	Metabolite	Bin	CRS [%]	CRS > 42.82%	Rep
Acetate	114	Singlet	NA	114	Lactate	398	90.09	✓	398
Alanine	351*	79.77	✓	356		400	78.49	✓	
	354*	77.51	✓		Propionate	141	83.75	✓	141
	356	55.79	✓			17	80.36	✓	
Formate	514	Singlet	NA	514		15	80.01	✓	
Glucose	284	87.41	✓	284		139	76.63	✓	
	368	87.30	✓			12	75.20	✓	
	370	87.13	✓			142	66.26	✓	
	303	86.41	✓		Threonine	313	18.17	×	-
	278	86.35	✓		Trehalose	362	70.75	✓	362
	439	86.32	✓			360*	67.07	✓	
	360*	86.18	✓			349*	66.61	✓	
	343	85.92	✓			347*	57.38	✓	
	341	85.75	✓			354*	45.27	✓	
	337	85.74	✓			291	42.94	✓	
	434	85.66	✓			438	35.35	×	
	301	85.53	✓		Tryptophan	262	31.91	×	-
	297	85.46	✓			397	29.19	×	
	351*	85.04	✓			300	20.56	×	
	352	83.53	✓			476*	8.23	×	
	361	81.85	✓			474*	6.57	×	
	290	81.32	✓		Tyrosine	474*	95.51	✓	466
	288	80.75	✓			466	95.49	✓	
	250	62.90	✓			228	95.42	✓	
	346*	56.47	✓			476*	95.36	✓	
Glutamate	150	36.59	×	-		226	95.16	✓	
	346*	13.77	×			475	94.47	✓	
	130	3.41	×			241	94.17	✓	
	350	-8.84	×			244	94.09	✓	
Glutamine	349*	9.41	×	-		224	93.99	✓	
	347*	9.07	×			221	93.62	✓	
	140	-0.28	×			473	92.34	✓	
Glycine	309	Singlet	NA	309		477	92.01	✓	
Isoleucine	4	87.01	✓	4	Valine	318	18.56	×	-

*: Overlapping bins

Post CRS selection, only eight out of 15 metabolites remained in the shortlist. Amongst the excluded metabolites were; glutamate, glutamine, threonine, tryptophan and valine. The remaining metabolites (Table 3.6-2) span metabolite classes of: amino acids, carboxylic acids, and saccharides.

Table 3.6-2: Selected representative metabolites for *Ae. aegypti* pupa, representing the differences between males and females.

Class	Metabolite	Representative bin	Chemical shift [ppm]	KEGG code
Amino acids	Alanine	356	3.80	C00041
	Glycine	309	3.57	C00037
	Isoleucine	4	0.94	C00407
	Tyrosine	466	6.90	C00082
Carboxylic acids	Acetate	114	1.92	C00033
	Formate	514	8.46	C00058
	Lactate	398	4.11	C00186
	Propionate	141	2.18	C00163
Saccharides	Glucose	284	3.42	C00031
	Trehalose	362	3.85	C01083

Prior to MSEA, data was filtered with the selected bins to represent the metabolites of interest and to establish the discriminatory properties of selected metabolites exclusively. PCA (Figure 3.6-4-A) was performed on the new dataset consisting only of selected metabolites. PCA scores of PC1 (60.24%) and PC2 (14.86%) showed the most variation representing the difference between males and females. PC1 and PC2 cumulatively explained 75.10% of the variance. A total of six components were required to explain the 95% of the variance in the data. A cross-validated supervised PLS-DA model was made to accentuate these differences (Figure 3.6-4-B).

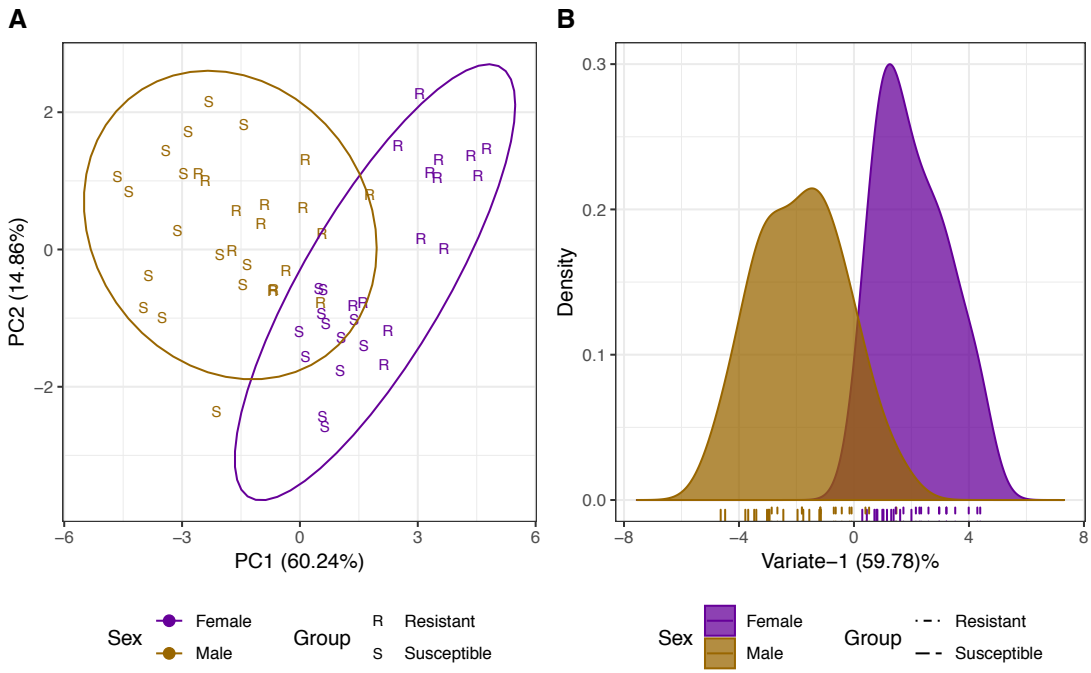


Figure 3.6-4: A) PCA on selected metabolite for wild type *Ae. aegypti* pupae, coloured by the sexes (n_{Female}=26, n_{Male}=29). A total of six components were required to explain the 95% variance in the data. Ellipses represent the 95% confidence region. B) PLS-DA density plot showing discrimination of females against males (n_{Female}=26, n_{Male}=29) in wild type *Ae. aegypti* pupae. Each tick represents a sample from groups. Variate-1 explains 59.78% of the variance in the data. The model accuracy was measured as 100%.

PLS-DA model performance was optimised to a single-variate model complexity with 100% accuracy (Appendix 11 for further metrics). The density plot shows the models capability of discriminating between males and females. To understand the metabolite level differences, a BH adjusted t-test method was performed.

To further understand the changes between resistant and susceptible pupae strains, metabolite levels were compared *via* BH adjusted t-test (for detailed statistics see Appendix 12). Compared metabolites were then visualised by boxplots (Figure 3.6-5). Out of the 10 selected metabolites, all were significantly different between male and female species. These 10 metabolites were then analysed by MSEA to compare the pathways affected by sex in pupae and adult mosquitoes.

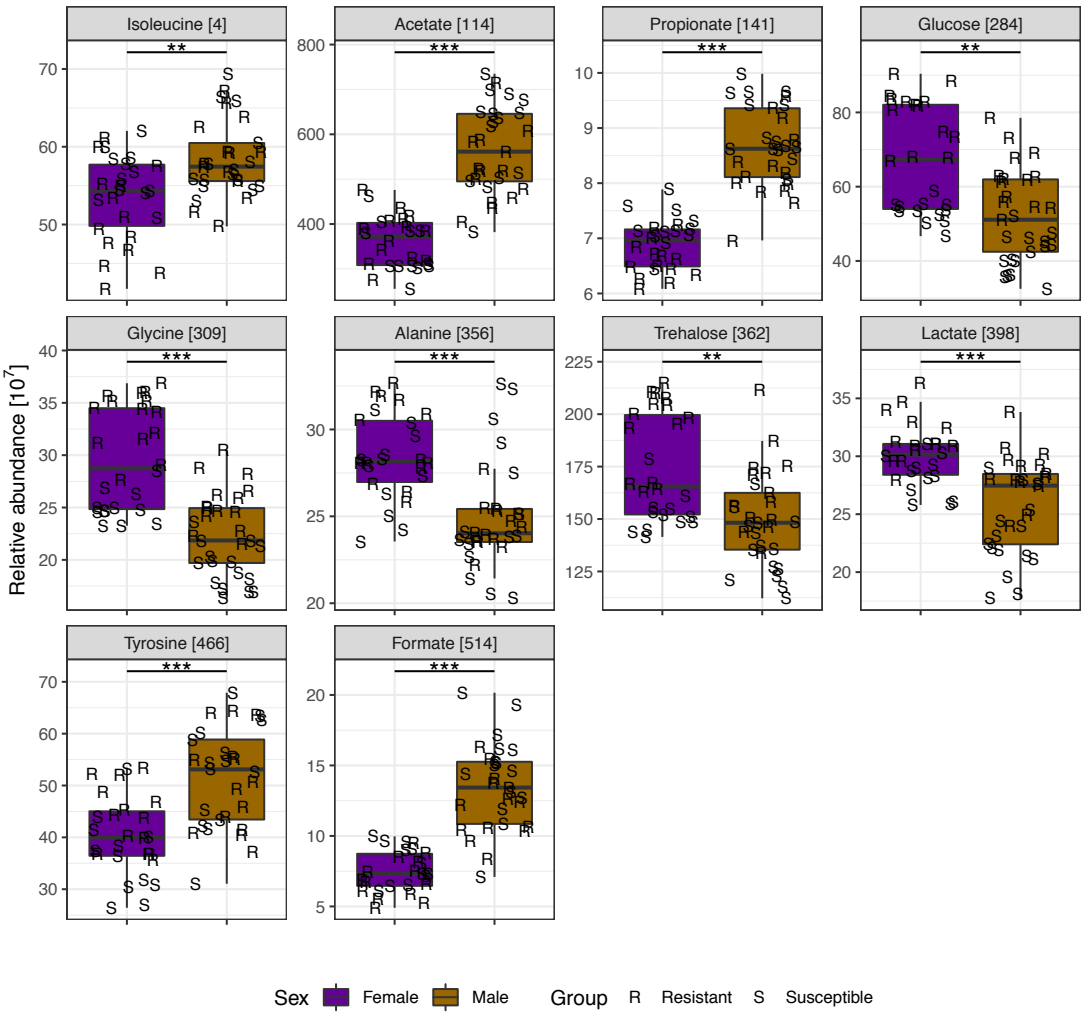


Figure 3.6-5: Boxplot of metabolites selected from *Ae. aegypti* pupae (n_{Female}=26, n_{Male}=29). Metabolites were compared *via* BH adjusted t-test. ** and *** denotes p-values less than 0.01 and 0.0001 respectively.

3.6.2 Sex-specific differences in adult mosquitoes

3.6.2.1 Statistical analysis

In order to obtain comparable results, the approach taken in pupal analysis was applied to the adult dataset. Prior to statistical analyses, batch variation in the data was estimated *via* PVCA. Batch variation contribution in the raw data was revealed to be 0.86%. When the dataset was PQN normalised and auto-scaled, batch contribution was increased to 3.34%. When compared to other batch correction methods, ComBat was able to correct the batch effectively by reducing its contribution to 0.00%. Prior to further analysis, data was PQN normalised, ComBat corrected and auto-scaled in the given order. Following the batch correction, variance in the data was observed *via* PCA.

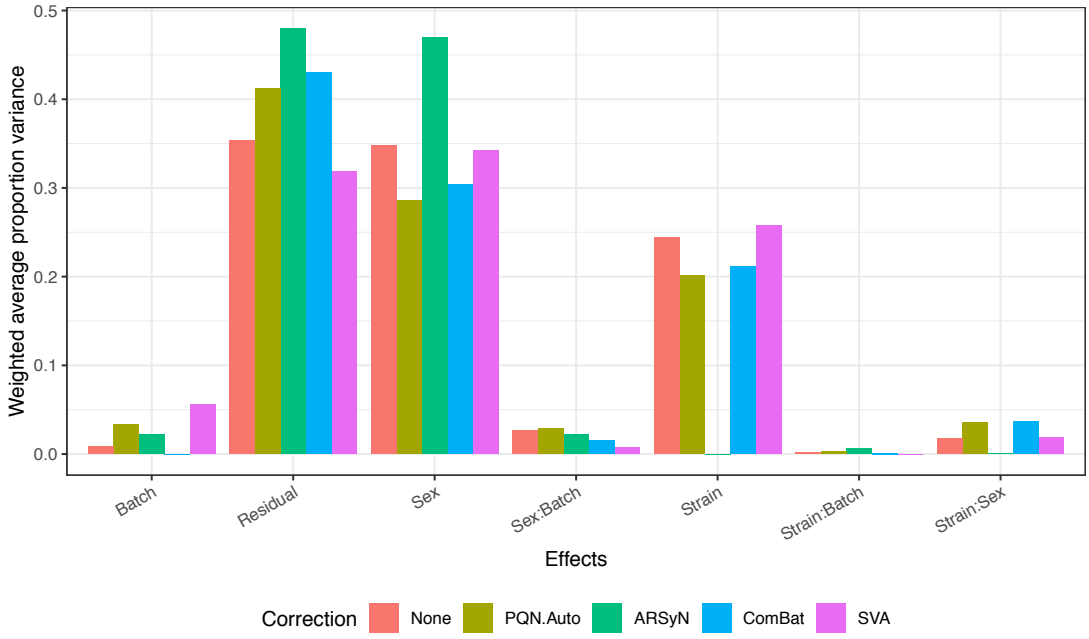


Figure 3.6-6: PVCA estimation of *Ae. aegypti* adult data (n_{Female}=22, n_{Male}=19). Each bar colour represents a different data treatment method. Each batch correction method was normalised before correction and scaled after with PQN normalisation and auto scaling respectively. Where effects of two factors were estimated ':' was used to show it. Residual, total remaining (residual) variation in the data; PQN.Auto, PQN normalisation and Auto scaling.

PCA (Figure 3.6-7-A) was performed to observe the major variances in the data. Variances representing the sexes were found along a diagonal of PC1 (29.14%) and PC2 (13.17%). PC1 and PC2 accounted for 42.31% of the variance in the data. In order to explain the 95% variance in the data, a total of 25 components were required. Overall, the plot shows a variance structure based on sex difference, some samples are not as clearly separated as the others. Similar to the pupae data a sub-population due to the resistance status can be seen. By accentuating differences between males and females with a PLS-DA model, metabolites responsible for these differences can be selected.

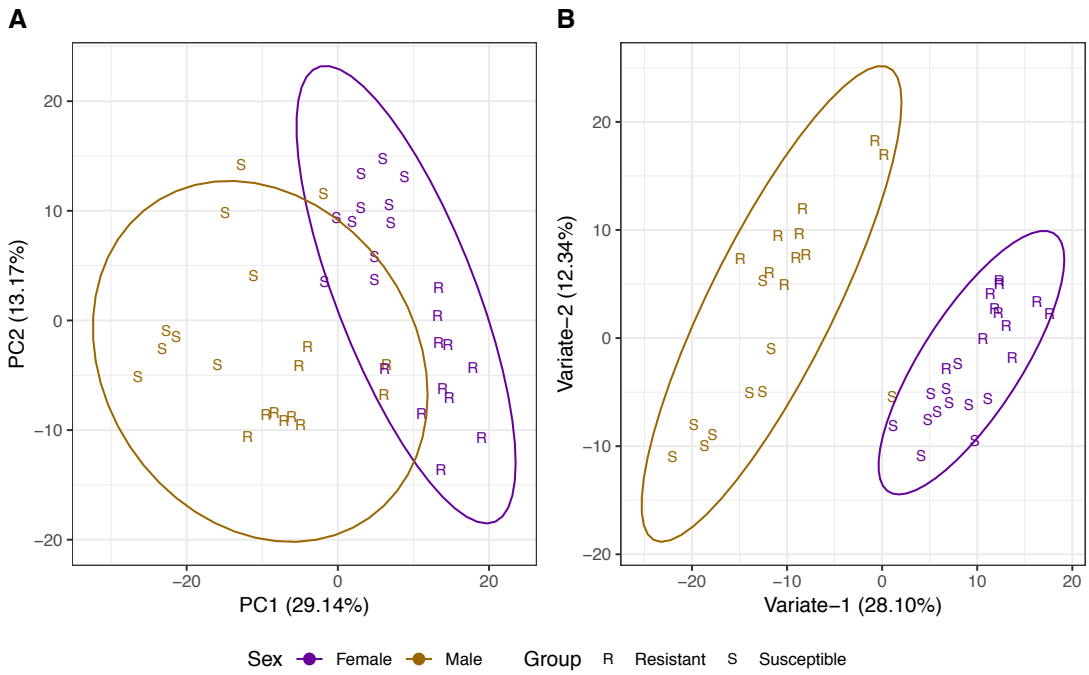


Figure 3.6-7: A) PCA for wild type *Ae. aegypti* adult showing PC1 (29.14%) against PC2 (13.17%) (n_{Female}=22, n_{Male}=19). PC1 and PC2 account for a cumulative explained variance of 42.31%. A total of 25 components were required to explain the 95% variance in the data. Ellipses represent 95% confidence region. B) PLS-DA scores plot of variate-1 (28.10%) against variate-2 (12.34%) discriminating between males and females of adult wild type *Ae. aegypti* (n_{Female}=22, n_{Male}=19). Model complexity of two-variate was determined through cross-validation with 100% accuracy. Brackets report the explained variance for the variate. Ellipses represent 95% confidence region for each group.

Discrimination between male and female samples were performed *via* PLS-DA modelling. Using a cross-validation, PLS-DA model was optimised to a two-variate model with 100% accuracy (Appendix 11 for further metrics). The PLS-DA (Figure 3.6-7-B) model could clearly separate males and females, except for one sample. This is a misclassification of the model, representing the limitations of the model on very extreme samples. VIP scores were calculated from the PLS-DA model in order to select metabolites for MSEA.

3.6.2.2 Key metabolites of the comparison

VIP scores were calculated for all the bins of the model. In order to create a list of selected metabolites, the most influential bins of the separation needed to be identified. To achieve this, only identified metabolites scoring higher than the threshold of one on variate-1 and variate-2 were considered (Figure 3.6-8). From the 513 VIP scores calculated, 196 bins scored above the threshold on both variate-1 and variate-2. Within the 196 bins, only 34 (17.35%) bins were identified representing 14 metabolites.

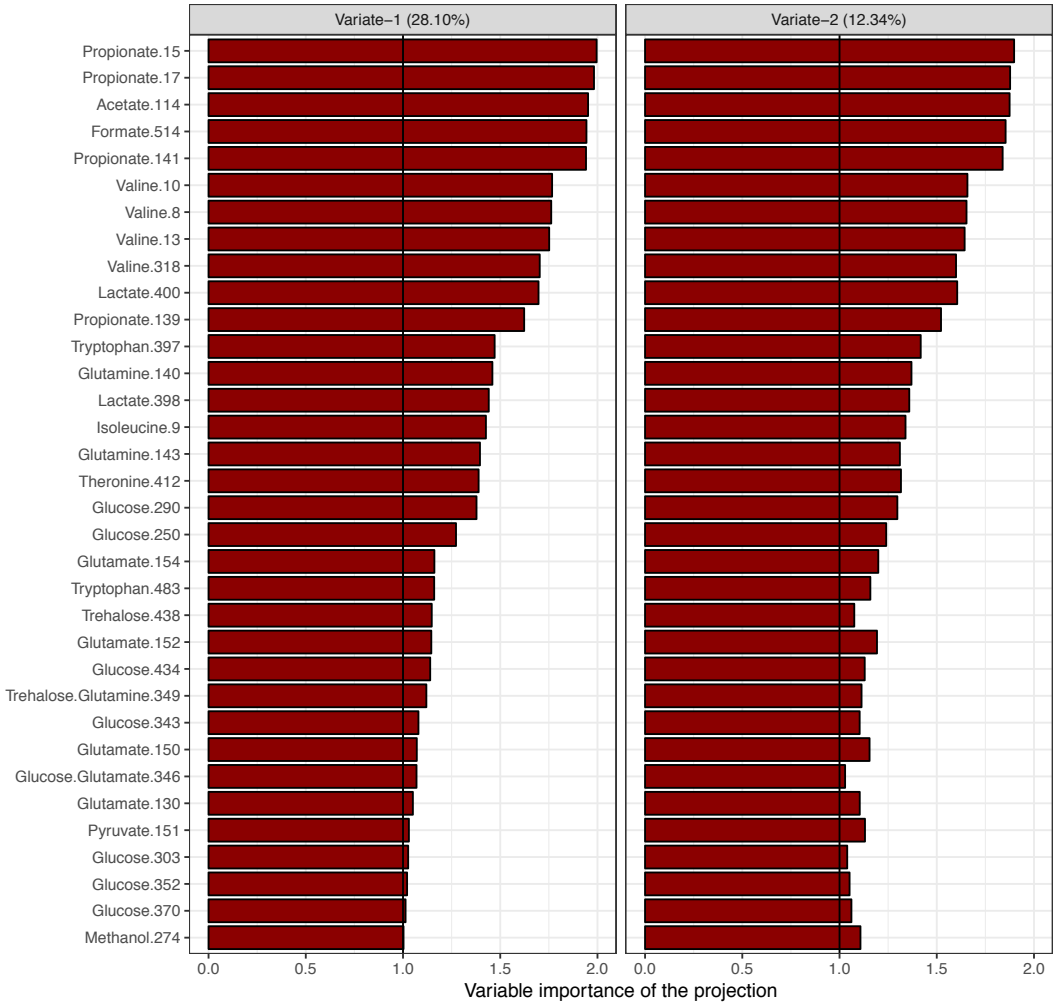


Figure 3.6-8: VIP scores for wild type adult *Ae. aegypti*. PLS-DA model discriminating between males and females. A total of 34 identified bins scored higher than 1, representing 14 metabolites. Black line represents VIP score of 1.

In order to generate a list of selected metabolites, a selection of bins representing each metabolite was required. Prior to this selection, bins were assessed on how well they represented each metabolite by CRS. A threshold was set to 41.68%, calculated from all the CRS. From the VIP-selected bins, the highest, non-overlapping (where applicable) scoring bins above the threshold were selected as the representative of its metabolite (Table 3.6-3).

Table 3.6-3: CRS for wild type *Ae. Aegypti* adults. A passing score of 41.68% was calculated using all identified bin CRS. Res: representative bin.

Metabolite	Bin	CRS [%]	CRS > 41.68%	Rep	Metabolite	Bin	CRS [%]	CRS > 41.68%	Rep
Acetate	114	Singlet	NA	114	Isoleucine	9	72.60	✓	9
Formate	514	Singlet	NA	514	Lactate	398	76.56	✓	398
Glucose	434	78.44	✓	434		400	62.34	✓	
	370	78.20	✓		Methanol	274	Singlet	NA	274
	303	78.00	✓		Propionate	15	77.02	✓	15

	343	77.45	✓			141	76.87	✓	
	352	69.21	✓			17	75.59	✓	
	250	67.47	✓			139	66.64	✓	
	290	61.42	✓		Pyruvate	151	Singlet	NA	151
	346*	57.16	✓		Threonine	412	23.29	×	-
Glutamate	150	33.86	×	-	Trehalose	438	47.62	✓	438
	152	32.90	×			349*	44.36	✓	
	154	31.45	×		Tryptophan	483	53.41	✓	483
	130	30.09	×			397	38.11	×	
	346*	1.24	×		Valine	8	68.31	✓	8
Glutamine	140	23.93	×	-		13	67.87	✓	
	349*	15.96	×			10	67.74	✓	
	143	-0.58	×			318	61.64	✓	

* Denotes overlapping bin.

Using the VIP scoring, 14 unique metabolites were listed. Post CRS, three metabolites were excluded from this shortlist due to low CRS. The excluded metabolites were glutamate, glutamine and threonine. The remaining n11 metabolites (Table 3.6-4) comprised of four metabolite classes; alcohols, amino acids, carboxylic acids and saccharides.

Table 3.6-4: List of selected metabolites from the PLS-DA model discriminating between sexes of adult *Ae. aegypti*.

Class	Metabolite	Representative bin	Chemical shift [ppm]	KEGG code
Alcohols	Methanol	274	3.36	C00132
Amino acids	Isoleucine	9	1.02	C00407
	Tryptophan	483	7.33	C00078
	Valine	8	1.00	C00183
Carboxylic acids	Acetate	114	1.92	C00033
	Formate	514	8.46	C00058
	Lactate	398	4.11	C00186
	Propionate	15	1.10	C00163
	Pyruvate	151	2.37	C00022
Sugars	Glucose	434	4.65	C00031
	Trehalose	438	5.20	C01083

Prior to MSEA, the discriminatory properties of these metabolites were exclusively probed using PCA (Figure 3.6-9-A). PCA scores plot of PC1 (60.19%) against PC2 (13.47%) show a clear separation between males and females. A wider spread in the male cluster suggests a larger variation in male population compared to the females. These two components account for a cumulative variance of 73.66%. A total of six components were required to explain 95% variance in the data. A cross validated PLS-DA model was also built in order to determine the performance of the selected metabolites in a statistical discriminatory method.

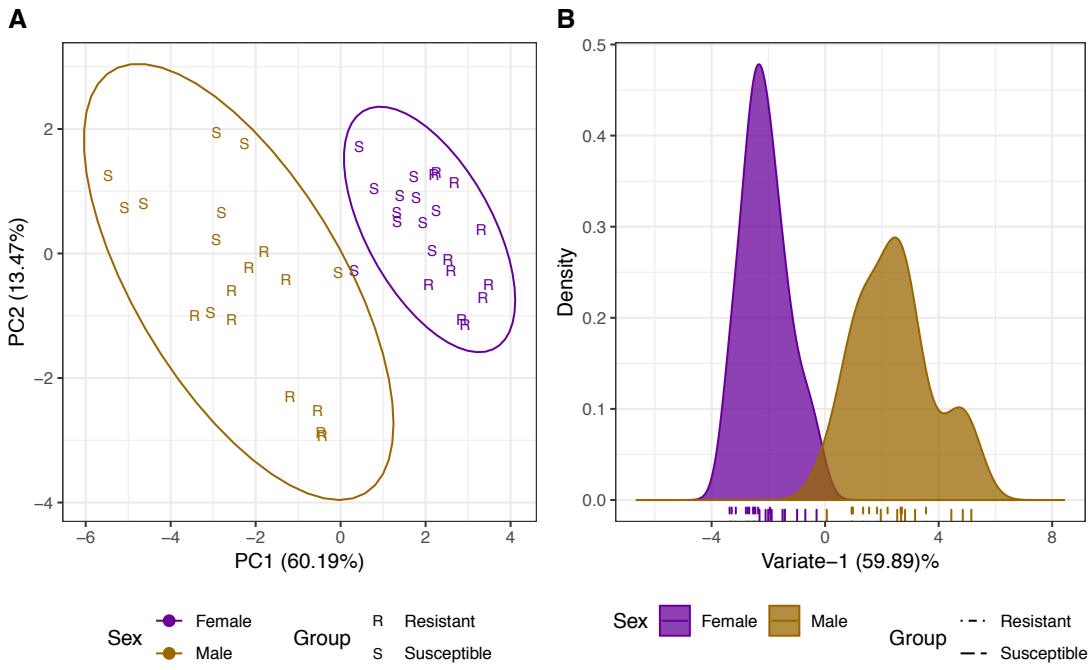


Figure 3.6-9: A) PCA of selected metabolites in adult wild type *Ae. aegypti* (n_{Female}=22, n_{Male}=19). PC1 (60.19%) and PC2 (13.47%) account for a cumulative variance of 73.66% while six components were required to explain the 95% variance in the data. Ellipses represent the 95% confidence region. B) PLS-DA density plot of adult wild type *Ae. aegypti* discriminated by sex (n_{Female}=22, n_{Male}=19). Single-variates model (59.89% explained variance) complexity was optimised using a cross-validation with 100% accuracy. Ticks represent samples from each group.

A cross-validated PLS-DA model was built in order to discriminate between males and females using the selected metabolites Figure 3.6-9-B. A single variate model with 100% accuracy (Appendix 11 for further metrics) was shown to be the optimal complexity for sex discrimination. Males could be discriminated from females on variate-1. Metabolite comparison was carried out in order to gain metabolite level insight on the changes between males and females.

Metabolite levels were shown *via* boxplot (Figure 3.6-10) and statistical significance was calculated using BH adjusted t-test (for detailed statistics see Appendix 12). All selected metabolites were found to be significantly different between males and females. More specifically, selected amino acids (valine, isoleucine and tryptophan) were all significantly higher in females, whereas all carboxylic acids (propionate, acetate, pyruvate and formate) except lactate were higher in males. This was also observed in both knock-down and wild type *An. gambiae*. In the saccharides, glucose was higher in females, meanwhile, trehalose was higher in males. Lastly, the only alcohol identified, methanol, was significantly higher in males.

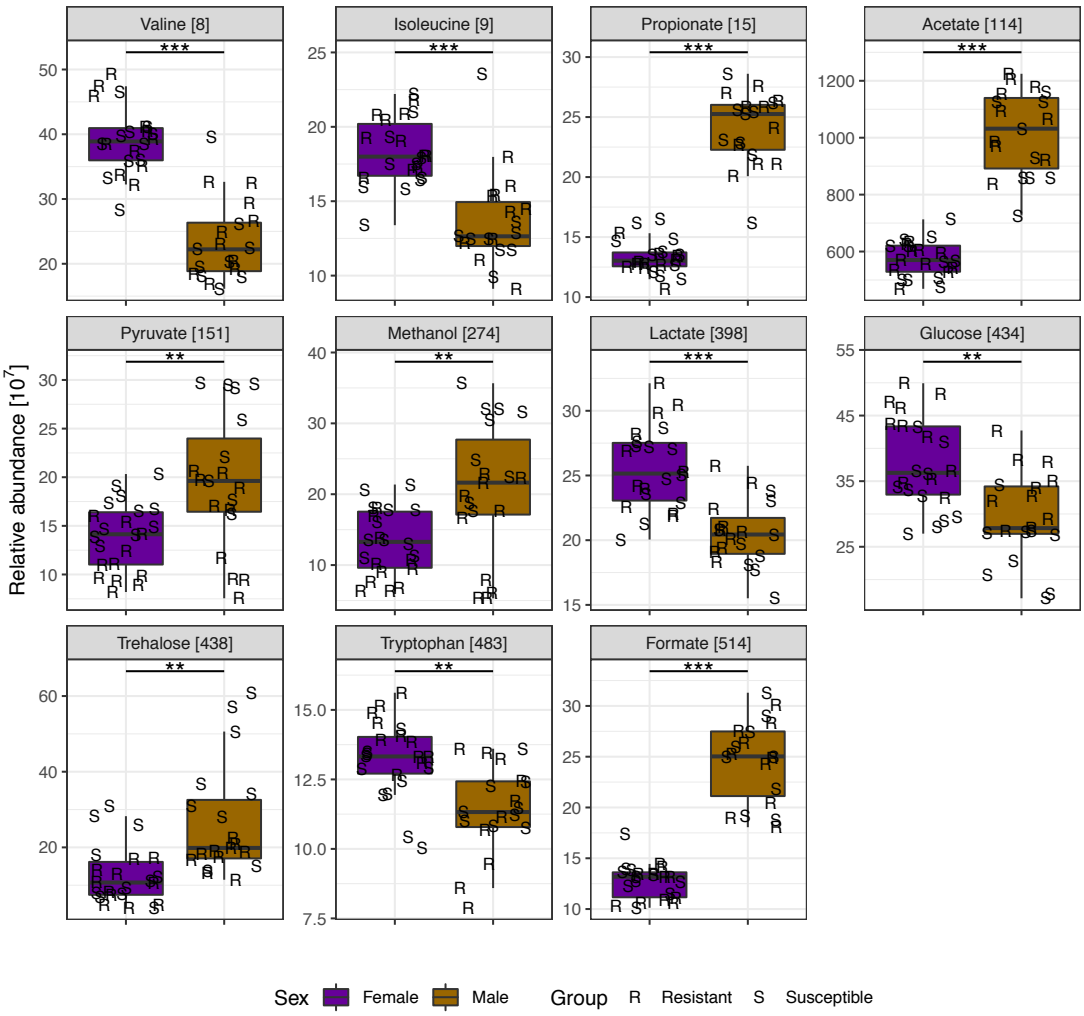


Figure 3.6-10: Boxplots of selected metabolites from PLS-DA model discriminating between females and males in adult wild type *Ae. aegypti* (n_{Female}=22, n_{Male}=19). ** and *** represents p-values less than 0.01 and 0.001 respectively.

3.6.3 Sex-specific differences across stages

Metabolite differences in pupae and adults were summarised in Table 3.6-5. The total number of metabolites selected from the pupae model was 10 compared to the 11 metabolites from the adult model. Overlapping carboxylic acids follow the same trend both in pupae and adults. In the amino acids, only isoleucine metabolite is shared and it lower in female pupae while higher in female adults. This could be due to the development of the pupae. From the sugars, only glucose was found to be higher in both adult and pupae females. For pathway level information, MSEA was performed on the selected metabolites.

Table 3.6-5: Metabolite changes summary for wild type *Ae. aegypti* pupae and adults. Arrows represent significant metabolite level differences when were females compared to males, and square brackets report BH-adjusted p-values.

		Pupa	Adult	
Class		**Female compared to male**		
Alcohols			↓ [4.61 x10^{-3}]	Methanol
Amino acids	Alanine	↑ [2.82x10^{-5}]		
	Glycine	↑ [4.38 x10^{-7}]		
	Isoleucine	↓ [1.59 x10^{-5}]	↑ [1.59 x10^{-5}]	Isoleucine
			↑ [5.92 x10^{-4}]	Tryptophan
	Tyrosine	↓ [2.53 x10^{-5}]		
			↑ [1.25 x10^{-9}]	Valine
Carboxylic acids	Acetate	↓ [1.05 x10^{-11}]	↓ [5.92 x10^{-11}]	Acetate
	Formate	↓ [2.75 x10^{-11}]	↓ [6.63 x10^{-11}]	Formate
	Lactate	↑ [2.53 x10^{-11}]	↑ [3.63 x10^{-6}]	Lactate
	Propionate	↓ [1.23 x10^{-11}]	↓ [3.28 x10^{-12}]	Propionate
			↓ [3.85 x10^{-3}]	Pyruvate
Sugars	Glucose	↑ [1.10 x10^{-4}]	↑ [5.92 x10^{-4}]	Glucose
	Trehalose	↑ [3.99 x10^{-4}]	↓ [1.17 x10^{-3}]	Trehalose

When MSEA (Figure 3.6-11) was performed on selected metabolites, seven pathways were significantly over-represented. Out of the seven, two were uniquely significant for adults and the remaining five were commonly significant (Table 3.6-6).

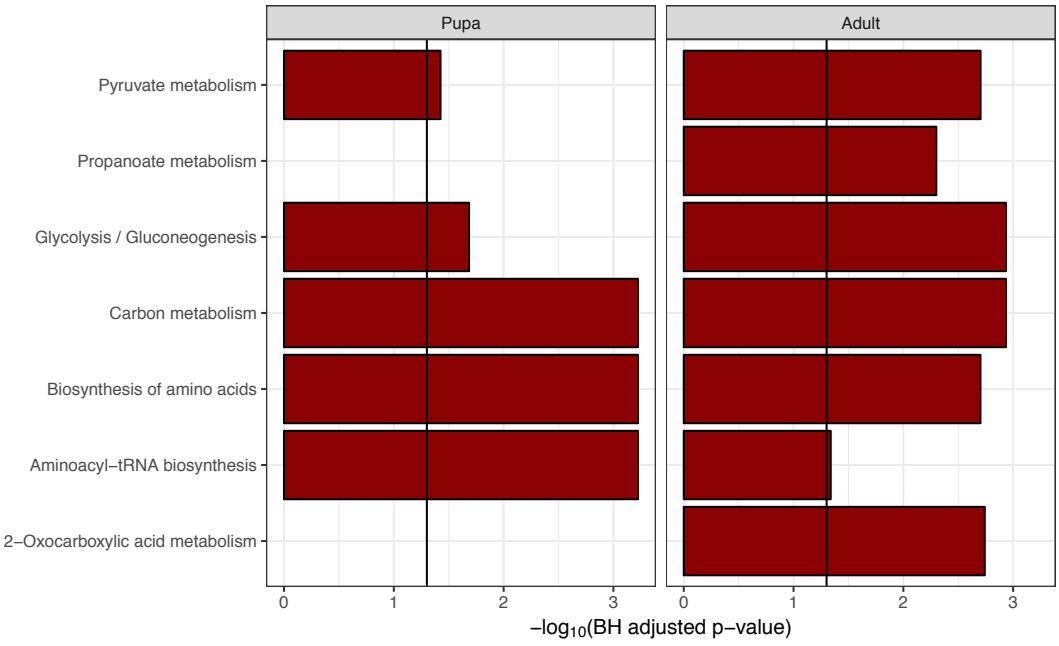


Figure 3.6-11: MSEA from selected metabolites in wild type *Ae. aegypti* pupae and adults. Black line represents p-value of 0.05.

Table 3.6-6: MSEA result details for wild type *Ae. aegypti*, reporting; stage, raw & BH adjusted p-values, number of hits and matched metabolites.

Pathway	Stage	Raw p-value	BH adjusted p-value	Metabolites	Hits/total (%)
2-Oxocarboxylic acid metabolism	Adult	0.00015	0.00181	4/134 (2.99%)	Pyruvate, Tryptophan, Valine, Isoleucine
Aminoacyl-tRNA biosynthesis	Pupa	0.00005	0.00060	4/52 (7.69%)	Glycine, Alanine, Tyrosine, Isoleucine
	Adult	0.00892	0.04586	3/52 (5.77%)	Tryptophan, Valine, Isoleucine
Biosynthesis of amino acids	Pupa	0.00003	0.00060	4/128 (3.13%)	Glycine, Alanine, Tyrosine, Isoleucine
	Adult	0.00028	0.00198	4/128 (3.13%)	Pyruvate, Tryptophan, Valine, Isoleucine
Carbon metabolism	Pupa	0.00002	0.00060	4/112 (3.57%)	Acetate, Glycine, Alanine, Formate
	Adult	0.00005	0.00116	4/112 (3.57%)	Pyruvate, Acetate, Formate, Methanol
Glycolysis / Gluconeogenesis	Pupa	0.00229	0.02063	3/31 (9.68%)	Glucose, Acetate, Lactate
	Adult	0.00006	0.00116	4/31 (12.9%)	Pyruvate, Glucose, Acetate, Lactate
Propanoate metabolism	Adult	0.00083	0.00501	4/48 (8.33%)	Acetate, Methanol, Propionate, Lactate
Pyruvate metabolism	Pupa	0.00521	0.03753	3/31 (9.68%)	Acetate, Formate, Lactate
	Adult	0.00022	0.00198	4/31 (12.9%)	Pyruvate, Acetate, Formate, Lactate

Following the analyses of sex-specific metabolic signatures of wild type strains (*An. gambiae* and *Ae. aegypti*), it is evident that the metabolite differences between sexes are greater in the wild type *Ae. aegypti* compared to the *An. gambiae* samples. In *Ae. aegypti* these differences were much heightened in pupae compared to adults. Two sub-populations can clearly be seen in *Ae. aegypti* pupae, whereas in adults these sub-population cannot be seen. Interestingly, when *Ae. aegypti* sex differences were analysed *via* supervised PLS-DA, both pupae and adult models required a higher number of metabolites. A lower number of metabolites was expected, as observed in *An. gambiae* analyses. When the analyses of pupal and adult sex differences are compared, there are five metabolites (acetate, formate, lactate, propionate and glucose) significantly different between males and females while their levels are similar across stages within each sex (Table 3.6-5). Additionally, between pupae and adults, the pathways pyruvate metabolism, glycolysis/gluconeogenesis, carbon metabolism, biosynthesis of amino acids, and aminoacyl-tRNA biosynthesis were significantly over-represented as pathways altering between males and females. Therefore, the precautions taken in further analyses to mitigate sex differences in the next chapters can be applied both to pupal and adult datasets enabling the increase of sample size.

3.7 Chapter results summary

This chapter aimed to explore the sex differences in the metabolome of knock-down *An. gambiae*, wild type *An. gambiae*, and wild type *Ae. aegypti*. Each dataset was analysed in two parts as pupae and adults while using both unsupervised and supervised multivariate methods. Sex differences were less pronounced in knock-down *An. gambiae* compared to wild type *An. gambiae* and *Ae. aegypti.* Pupae separations were more complex than adults, as revealed by the PLS-DA model scores (Table 3.7-1). The wild type *Ae. aegypti* pupal model (selected metabolites) and adult models (all and selected metabolites) were the most robust (100% accuracy), whereas knock-down *An. gambiae* pupa model for all data was the poorest (45.45% accuracy). Overall, adult models were more robust in discriminating between males and females.

Table 3.7-1: PLS-DA model details for discrimination between males and females.

Species	Genotype	Stage	Data coverage	Model variates	Accuracy	VIP	Metabolites selected
An. gambiae	Knock-down	Pupa	All data	3	45.45%	24 (23.30%)	9
		Adult		2	94.12%	20 (18.35%)	6
		Pupa	Selected	1	63.64%	NA	NA
		Adult		2	94.12%	NA	NA
	Wild type	Pupa	All data	2	71.43%	53 (31.26%)	12
		Adult		2	87.50%	58 (32.40%)	8
		Pupa	Selected	2	71.43%	NA	NA
		Adult		2	93.75%	NA	NA
Ae. aegypti	Wild type	Pupa	All data	1	93.75%	60 (28.71%)	10
		Adult		2	100.00%	34 (17.35%)	11
		Pupa	Selected	1	100.00%	NA	NA
		Adult		1	100.00%	NA	NA

Metabolite selection was carried out from the cross-validated PLS-DA models and metabolite levels were compared using BH-adjusted t-test. Table 3.7-2 collates all the changes across species and stages. The changes shown in the table are most concentrated in the metabolites belonging to the carboxylic acids class (acetate, formate, lactate, and propionate). Out of all the selected metabolites, carboxylic acids are the most consistent in terms of metabolite level between males and females.

Table 3.7-2: Metabolite level comparisons between sexes for knock-down *An. gambiae*, wild type *An. gambiae* and wild type *Ae. aegypti*. Arrows represent significant changes, NS (arrow) denotes non-significant change with mean abundance level, square brackets report the BH-adjusted p-value, and '-' represents metabolite not selected from the PLS-DA model.

		Knock-down *An. gambiae*		**Wild type *An. gambiae***		**Wild type *Ae. aegypti***	
		Females compared to males					
Class	**Metabolite**	**Pupa**	**Adult**	**Pupa**	**Adult**	**Pupa**	**Adult**
Alcohols	Methanol	NS (↑) [3.12×10^{-1}]	-	-	-	-	↓ [4.61×10^{-3}]
Amino acids	Alanine	-	-	↑ [4.00×10^{-5}]	↑ [2.74×10^{-3}]	↑ [2.82×10^{-5}]	-
	Glutamate	-	-	↓ [1.82×10^{-3}]	-	-	-
	Glutamine	-	-	↑ [1.09×10^{-2}]	-	-	-
	Glycine	NS (↓) [3.12×10^{-1}]	-	-	↑ [4.38×10^{-3}]	↑ [4.38×10^{-7}]	-
	Isoleucine	-	-	-	-	↓ [1.59×10^{-5}]	↑ [1.59×10^{-5}]
	Tryptophan	NS (↓) [3.93×10^{-1}]	-	-	-	-	↑ [5.92×10^{-4}]
	Tyrosine	NS (↑) [8.93×10^{-1}]	↓ [4.14×10^{-6}]	-	-	↓ [2.53×10^{-5}]	-
	Valine	-	-	↑ [3.43×10^{-4}]	-	-	↑ [1.25×10^{-9}]
Carboxylic acids	Acetate	NS (↓) [2.43×10^{-1}]	-	↓ [1.82×10^{-3}]	↓ [3.04×10^{-5}]	↓ [1.05×10^{-11}]	↓ [5.92×10^{-11}]
	Formate	NS (↓) [3.12×10^{-1}]	-	↓ [2.09×10^{-3}]		↓ [2.75×10^{-11}]	↓ [6.63×10^{-11}]
	Fumarate	-	-	↓ [3.57×10^{-2}]	-	-	-
	Lactate	-	↑ [1.25×10^{-21}]	↑ [3.43×10^{-4}]	↑ [3.63×10^{-10}]	↑ [2.53×10^{-11}]	↑ [3.63×10^{-6}]
	Propionate	NS (↓) [3.61×10^{-1}]	↓ [7.98×10^{-5}]	↓ [2.53×10^{-4}]	↓ [1.08×10^{-3}]	↓ [1.23×10^{-11}]	↓ [3.28×10^{-12}]
	Pyruvate	-	-	↓ [2.17×10^{-2}]	-	-	↓ [3.85×10^{-3}]
Purines	Oxypurinol	NS (↓) [8.93×10^{-1}]	↓ [1.23×10^{-6}]		↓ [3.46×10^{-6}]	-	-
	Xanthine	-	↑ [3.12×10^{-1}]	-	-	-	-
Saccharides	Glucose	NS (↑) [8.93×10^{-1}]	-	↑ [4.00×10^{-5}]	↑ [4.38×10^{-3}]	↑ [1.10×10^{-4}]	↑ [5.92×10^{-4}]
	Trehalose	-	↓ [1.30×10^{-3}]	↑ [2.82×10^{-3}]	↑ [4.38×10^{-3}]	↑ [3.99×10^{-4}]	↓ [1.17×10^{-3}]

Over-represented pathway comparisons are summarised in Table 3.7-3. In knock-down *An. gambiae* pupae, carbon metabolism and propanoate metabolism were found to be the only over-represented pathways between males and females. The wild type *An. gambiae* comparison identified the highest number of pathways with 15 pathways. Amongst the 15 pathways, alanine, aspartate and glutamate metabolism, aminoacyl-tRNA biosynthesis, arginine biosynthesis, biosynthesis of amino acids, butanoate metabolism, glyoxylate and dicarboxylate metabolism, nicotinate and nicotinamide metabolism and nitrogen metabolism were specific to pupae. Meanwhile, only ABC transporters, carbon metabolism and glycolysis/gluconeogenesis were common between pupae and adults. Only propanoate metabolism was found differing in adults. In wild type *Ae. aegypti*, common pathways to both

pupae and adults were aminoacyl-tRNA biosynthesis, biosynthesis of amino acids, carbon metabolism, glycolysis/gluconeogenesis, and pyruvate metabolism. Meanwhile, 2-oxocarboxylic acid metabolism and propanoate metabolism were identified as differing in adults.

Table 3.7-3: MSEA summary for sex comparison between different species and strains of mosquitoes.

Over-represented pathways	**Knock-down** ***An. gambiae***	**Wild type** ***An. gambiae***	**Wild type** ***Ae. aegypti***
2-oxocarboxylic acid metabolism		P	A
ABC transporters		C	
Alanine, aspartate and glutamate metabolism		P	
Aminoacyl-tRNA biosynthesis		P	C
Arginine biosynthesis		P	
Biosynthesis of amino acids		P	C
Butanoate metabolism		P	
Carbon metabolism	P	C	C
Glycolysis/Gluconeogenesis		C	C
Glyoxylate and dicarboxylate metabolism		P	
Nicotinate and nicotinamide metabolism		P	
Nitrogen metabolism		P	
Propanoate metabolism	P	A	A
Pyruvate metabolism		P	C
Taurine and hypotaurine metabolism		P	

A: only in adults, C: common for pupae and adults, P: only in pupae.

3.8 Chapter Discussion

The work presented in this chapter evaluated the metabolic differences between male and female pupal and adult stage mosquitoes. NMR metabolomics highlighted sex differences that could be attributed to an energy-related mechanism due to the female mosquitoes' role in reproduction. In order to complete oviposition, female mosquitoes require a blood meal which may necessitate longer flight time in order to find a suitable source. Hence, females may be expected to be able to process and store energy more efficiently than males. Both *An. gambiae* (knock-down and wild type) and *Ae. aegypti* presented different degrees of metabolite differences between male and female at both pupa and adult stages. *Ae. aegypti* strains exhibited a more distinct metabolic profile between males and females than *An. gambiae*; this difference could be observed both in pupal and adult datasets *via* unsupervised methods (i.e. PCA). Additionally, sub-populations were observed both in *An. gambiae* and *Ae. aegypti*. These were attributed to their resistance status. Interestingly, wild type *An. gambiae* (pupae and adults) demonstrated more subtle distinctions between metabolite

profiles of males and females, yet they were still distinguishable *via* PCA. Surprisingly, in spite of observable physical differences, the distinction between males and females was less clear in the knock-down mosquitoes of *An. gambiae*. This suggests that the effects of the knock-down are not just to the CHC profile but also in the general metabolic characteristics of the mosquitoes. When males and females were compared, the CONT groups were also included. CONT did not form a sub-population or show any clustering in terms of sex in the PCA scores plot (Appendix 14). The commonality between the KD16, KD17 and CONT is the presences of the Gal4 driver, suggesting that the Gal4 driver expression might be affecting the sex characteristics observable *via* NMR metabolomics.

In terms of batch variation, knock-down pupae showed the most variation. Although it could be thought that this is the output of poorer signal to noise ratio (SNR), sample verification showed that there were no significant SNR variations between any of the samples (Appendix 23) and that no group had a distinctly different SNR (either pupae or adults) which could explain the batch variation observed. One explanation could be linked to one of the caveats of the Gal4/UAS knock-down system: the levels of knock-down varies between individuals depending on the Gal4 expression. Also, due to the critical nature of these knock-downs to survival, it is thought that only the low-level knock-down pupae emerge and survive as adults. Hence, adults bearing the knock-down are more likely to be similar between batches. In contrast, during the pupal stage, all different levels of knock-downs are alive during sample collection (with no selection based on fluorescence possible), thus increasing the likelihood of a higher degree of batch variation in pupae samples.

In this chapter, sex-related differences were investigated, with a particular focus on understanding how strong these differences are compared to the experimental differences. This was accomplished by gathering all samples of a single experiment, including controls, be it comparison of resistant and susceptible strains or knock-down models and dividing them into males and females. The main goal in doing so was to answer the question 'Would sex-related differences be so strong that it would mask all the experimental differences?' During this process, groupings on experimental conditions were used as visual cues, but not in statistical analyses. Alternatively, this could have been done with smaller groups such as males vs females in only KD16 knock-downs or resistant and susceptible strains of *Ae. aegypti*. This would have resulted in very specific sex-related differences, if they were observed, and considering the number of comparisons, joining these results would have been more challenging. The approach that was taken in this project was to show a more

generalised result for the sex-differences. The expected confounding differences that would have manifested themselves in the PCA scores plots were the experimental groups (i.e. knock-down status or resistance status) which were visually highlighted.

Feeding habit is a factor that could not be fully controlled in this study, although food source was consistent throughout. All adult mosquitoes were supplied with a 10% sucrose solution as food source, however the amount ingested and timing relative to collection could not be controlled. The PCA scores plot of wild type adult *An. gambiae* exhibited a sub-population in females. When further investigated *via* PCA, this sub-population was heavily influenced by glucose (Appendix 13), a strong indication that the nutritional state of some mosquitoes was captured in the metabolomics of these samples.

Similar sub-population separation can also be observed in wild type of *Ae. aegypti*, although the sub-populations exhibited in both males and females were attributable to resistance status from the metadata. Differences between males and females were most pronounced in *Ae. aegypti* species compared to *An. gambiae*, as revealed by the separation in the PCA scores plot as well as the higher number of significantly different metabolites. This is thought to be mostly due to the females being larger than males. Ideally males and females would be normalised to dry body mass. Unfortunately, the drying and weighing process would have affected the metabolome. It is possible arrest the metabolic state with liquid nitrogen and dry the mosquitoes with lyophilisation. Although, from previous experience this application causes the lyophilised mosquito to be brittle and fragile making the handling of the mosquito challenging and therefore this step was skipped. In order to control for this source of variance, equal numbers of males and females were maintained in all experiments throughout the study and data was normalised using probabilistic quotient normalisation (PQN).

When PLS-DA models were used to discriminate between males and females, the key metabolites identified across species and stages that were attributable to sex include; glucose, trehalose, propionate, lactate, acetate. These metabolites are mainly involved in energy and energy storage mechanisms. Glucose is the main energy source [210] and it is converted into trehalose for short-term storage [210]. Trehalose reserves are kept constant and any excess is stored as glycogen [210]. Furthermore, the metabolites propionate, lactate and acetate are key metabolites found in pathways downstream of glycolysis such as TCA cycle and propionate metabolism. TCA continues the energy production mechanism started

by glycolysis. On the other hand propionate metabolism is an important pathway in synthesising the precursors for fatty acid biosynthesis such as malonyl-CoA and methyl-malonyl-CoA [78], [90]. Observing the over-representation of these pathways further supports the hypothesis that the main difference demonstrated here *via* metabolic profiles originates from the higher energy demand of females compared to the males in *An. gambiae* and *Ae. aegypti*. Over representation of propionate metabolism, and the similar male and female profiles in KD mosquito strains suggest that loss of observable sex characteristics in the metabolome could be a consequence of disturbance of normal fatty acid pathway caused by knocking-down of enzymes critical in HC production. Although, when the CONT group was investigated *via* PCA for sex difference, males and females could not be separated (Appendix 14). This suggests the more plausible explanation is the common presence of the Gal4 driver line effecting the observable sex characteristics.

In this chapter, sex differences to varying degrees in different species (*An. gambiae* & *Ae. aegypti*) and strains (KD16, KD17, CONT, resistant & susceptible) of mosquitoes were shown. For knock-down *An. gambiae*, these differences were found to be the least pronounced compared to the wild type mosquitoes, as shown by lesser separation observed *via* PCA and the lowest number of significantly different metabolites out of all comparisons (pupa: 0 and adult: 6). Comparing the results obtained from each species' pupa and adult analyses, metabolic profiles were similar in each species in terms of over-represented pathways and selected metabolites. Furthermore, the limited sex difference in knock-downs may be traced to the pathways affected by the reduction of Cyp4g16 & Cyp4g17. Due to the overall differences observed between sexes, the following experiments were designed with a balanced 50-50% sex separation to avoid unwanted biases.

The study explained in this chapter is a novel work that compares the male and female metabolic profiles of pupal and adult mosquitoes. Through statistical methods, metabolic differences between sexes has been identified to glucose, trehalose, lactate, acetate and propionate. Furthermore, evidence has been shown suggesting that knocking-down Cyp4g16 and Cyp4g17 can indirectly affect the fatty acid biosynthesis.

Chapter 4

4 Analysis of Cyp4g16 and Cyp4g17 knock-downs in *An. gambiae*

4.1 Introduction, chapter aims & objectives

The rise in insecticide resistance requires the exploration of alternative methods to keep vector borne diseases at bay. Thus, investigation of cuticular resistance is of great importance. Several studies have reported the high abundance of cuticular hydrocarbons (CHCs) in resistant species. The studies further reported the particularly high abundances of alkanes and methyl-branched alkanes in the CHC profiles of resistant species. The CHC biosynthesis pathway is relatively unexplored compared to the depth and breadth of knowledge that has been amassed for other known pathways (e.g. glycolysis, citrate cycle). The majority of the information on the CHC biosynthesis pathway is a culmination of information gathered from a variety of insects, obtained through labelled tracer studies [78], [171]. Although this approach can show the conversion of metabolites, the complete picture requires the identification and verification of the enzymes that carry out these conversions.

As explained in section 1.4.3, alkane and methyl branched CHC biosynthesis (Figure 4.1-1) begins with the precursors acetyl-CoA, valine and leucine. Acetyl-CoA is elongated using malonyl-CoA for alkanes. Valine and leucine are elongated with malonyl-CoA and methylmalonyl-CoA if more methyl branching is needed. These elongation steps are catalysed by an elongase and at each step the HC is attached to a CoA. These CoAs are cleaved by a reductase leaving an aldehyde group. The terminal carbonyl needs to be removed in order to produce the final product hydrocarbon (HC) which then gets transported to the cuticular layer *via* lipophorins [96]. This final decarbonylation step is catalysed by a CYP decarbonylase.

Studies conducted on *Drosophila melanogaster* revealed two *An. gambiae* cytochrome p450 (CYP) enzymes (Cyp4g16 and Cyp4g17) catalysing the decarbonylation step in CHC biosynthesis [75]. Balabanidou *et al* further found that these enzymes are highly abundant in insecticide resistant *An. gambiae* strains [75]. Furthermore, they are highly expressed in the abdomen where oenocytes (hydrocarbon synthesising cells) reside [75]. Previous studies have shown that oenocytes express high quantities of cytochrome P450 reductase (CPR), a redox partner which CYP enzymes require [211].

To verify the decarbonylase activity of Cyp4g16 and Cyp4g17, Balabanidou *et al* expressed both enzymes as fusion proteins with CPR. While Cyp4g16-CPR was successfully expressed, Cyp4g17-CPR could not be expressed at similar concentrations. Using an *in vitro* assay, Cyp4g16-CPR was successfully shown to convert *n*-octacosane aldehyde to *n*-heptacosane revealing its function as a decarbonylase. As a result of the low expression of Cyp4g17-CPR, the activity could not be validated. Due to the high sequence identity, as well as the similar expression profiles of Cyp4g17 to Cyp4g16, it is hypothesized that the two enzymes catalyse the same reaction (personal conversation with Dr Gareth Lycett). Even though high expression of Cyp4g17-CPR could not be achieved, a knock-down model could be utilised to show evidence of its function by correlating to Cyp4g16 activity.

By using a Gal4/UAS knock-down of the Cyp4g16 or Cyp4g17, metabolic differences related to precursor metabolites are expected to be determined. More specifically, it is hypothesised that an accumulation of precursor metabolites (valine, leucine, isoleucine and acetate [as a proxy of acetyl-CoA]) in the KD insects compared to the Gal4 controls will be observed due to the impaired ability of the knockdowns to synthesise CHCs.

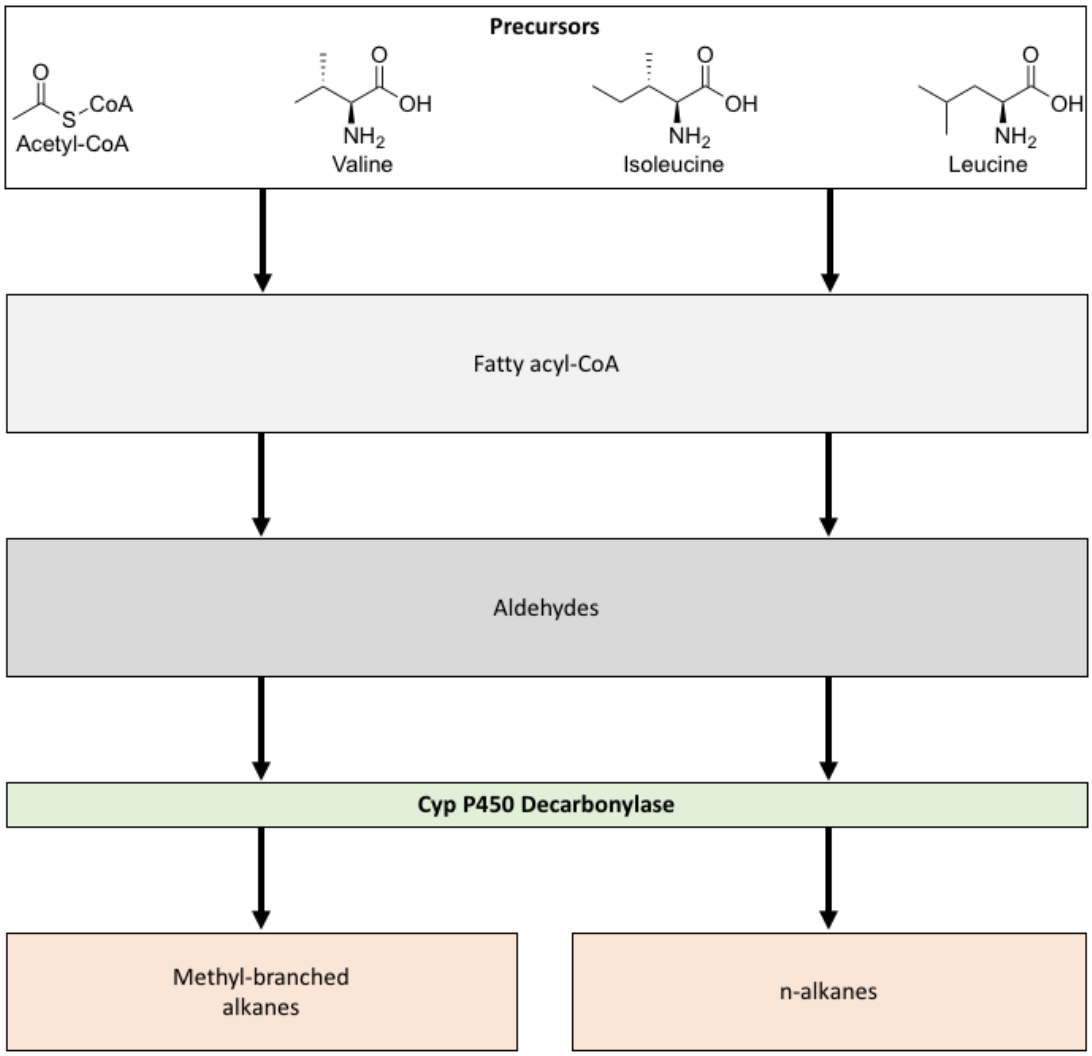


Figure 4.1-1: Simplified schematic of cuticular hydrocarbon production through Cyp P450 decarbonylases. Precursors (white box) are converted in to fatty acyl-CoA through a series of elongation and desaturation reactions. Fatty acyl-CoAs (light grey box) are then reduced to aldehydes (dark grey box) during which the CoAs are cleaved. Decarbonylation of aldehydes is facilitated by Cyp P450 decarbonylase (green box), yielding the final products (orange boxes).

4.2 Experimental design

In this chapter, Cyp4g16 and Cyp4g17 knock-down (KD16 and KD17) lines of *An. gambiae* were generated using a Gal4/UAS system by Dr Gareth Lycett of LSTM [188]. Gal4 homozygous progenies produces during breeding were used as controls (CONT) to the knock-downs strains. During the breeding of knock-down mosquitoes, the effects of the enzymes could be observed clearly, as shown in Figure 4.2-1. When control pupa bowls were compared to knock-down bowls, dead pupa and failed adult emergence where not more than 10%. The Cyp4g16 knock-down pupa bowl has a higher ratio of dead early pupa (approximately 60%) which sink to the bottom of the water, compared to Cyp4g17 and controls. On the other hand, the Cyp4g17 pupae bowl has a higher ratio of old pupa/eclosing

adults stuck in case and adults failing to emerge (approximately 90%) compared to Cyp4g16 knock-downs and controls. Mortality rates were found to be consistent with literature [188].

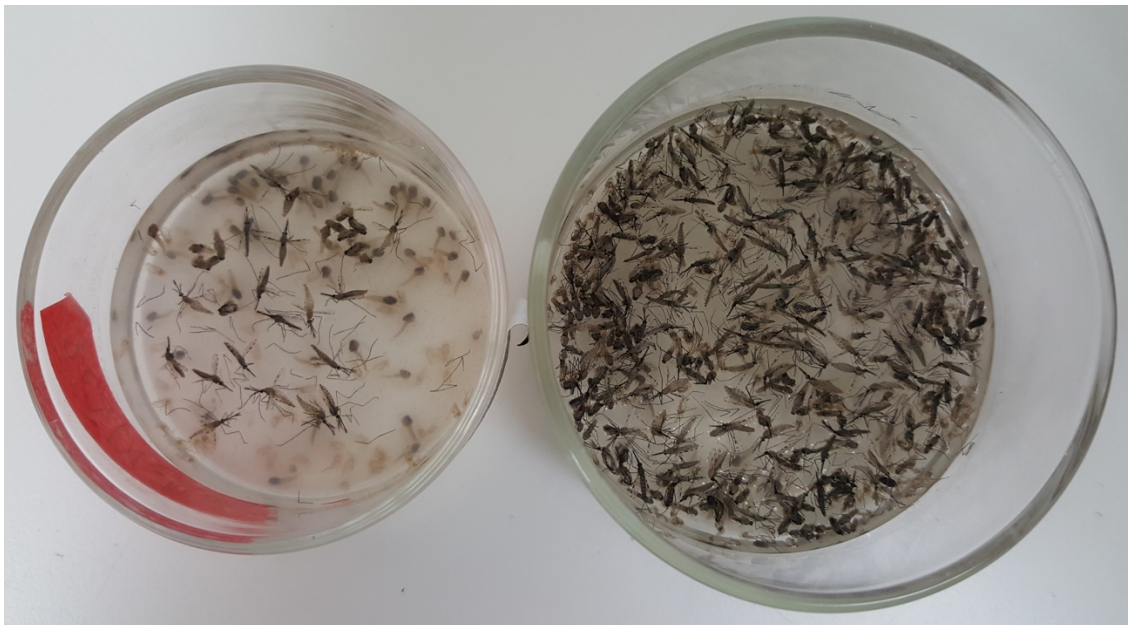


Figure 4.2-1: Phenotypic effects of Cyp4g16 (left) and Cyp4g17 (right) knock-downs from one breeding. In this breeding, Cyp4g16 knock-downs have high mortality rates during the pupal stage at 52% whereas Cyp4g17 knock-downs have extreme mortality rates in emergence at 92%. In following breeding cycles mortality rates were assessed visually (approximately 60% for KD16 pupae, 90% for KD17 adults and 10% for pupa and adult CONT) and were consistent with literature [188].

In order to collect the required number of samples for NMR metabolomics for polar metabolites, mosquito breeding was scaled up. Table 4.2-1 shows the sample counts for the data analysis per group in this chapter. Batch correction was applied as explained in Chapter 3.

Table 4.2-1: Sample Numbers used in KD16, KD17 and control (CONT) comparison.

	Collected samples			**Spectra acquired**			**Quality control passed**		
	KD16	**KD17**	**CONT**	**KD16**	**KD17**	**CONT**	**KD16**	**KD17**	**CONT**
Pupa	30	30	30	17	20	20	11	15	12
Adult	30	30	30	30	30	60	27	29	56

4.3 Metabolite assignment

Bin tables for the pupal and adult stages were identical and consisted of a total of 496 bins. Post identification, 114 bins were assigned (22.98%) to 21 unique metabolites (Table 4.3-1). Amongst the identified bins, overlapping bins totalled up to 15 (13.16%). Unidentified bins accounted for 77.02% with 382 bins. Metabolite assignment confidence were scored using an adapted version of the metabolite standards initiative's (MSI) scoring system. As a result

of this adapted MSI scoring, 16 metabolites were identified at Level 1 and five metabolites were identified at Level 2. Representative ^{1}H-NMR spectra for *An. gambiae* knock-downs grouped by experimental groups can be found in Appendix 15 (pupa) and Appendix 16 (adult).

Table 4.3-1: Metabolite assignment table, with MSI level, KEGG compound code and classifications.

Classification	Metabolite	Metabolite Identification Level (MSI)	Unique	Overlap	Total	KEGG code
Alcohols	Methanol	Level 1	1	0	1	C00132
Amino acids	Alanine	Level 1	2	3	5	C00041
	Glutamate	Level 1	7	3	10	C00025
	Glutamine	Level 1	2	3	5	C00064
	Glycine	Level 1	1	0	1	C00037
	Isoleucine	Level 1	3	0	3	C00407
	Threonine	Level 1	4	0	4	C00188
	Tryptophan	Level 1	12	3	15	C00078
	Tyrosine	Level 1	12	2	14	C00082
	Valine	Level 1	5	0	5	C00183
Carboxylic acids	Acetate	Level 1	1	0	1	C00033
	Formate	Level 2a	1	0	1	C00058
	Fumarate	Level 2a	1	0	1	C00122
	Lactate	Level 1	4	0	4	C00186
	Propionate	Level 2b	6	0	6	C00163
	Pyruvate	Level 1	1	0	1	C00022
	Succinate	Level 1	1	0	1	C00042
Purines	Oxypurinol	Level 2b	1	0	1	C07599
	Xanthine	Level 2b	1	0	1	C00385
Saccharides	Glucose	Level 1	24	8	32	C00031
	Trehalose	Level 1	10	7	17	C01083

4.4 Analysis of metabolic profiles of knock-downs

It is hypothesised that Cyp4g16 and Cyp4g17 catalyse the same reaction, although their location in the oenocytes are different [75]. Additionally, it is hypothesised that Cyp4g16 enzymes are active earlier in mosquito development, in the early larval stages, and remain active throughout adulthood, whereas Cyp4g17 transcription dramatically increases in 4th instar larvae and remains expressed throughout the later stages of life. Hence, it is expected that differences in KD16, KD17 and CONT. More specifically, KD16 is expected to exhibit more pronounced differences in pupae compared KD17 and CONT. Additionally a higher degree of similarity between KD16 and KD17 is expected in the adult stage but different to CONT.

4.4.1 Analysis of pupae metabolic profile

4.4.1.1 Statistical analysis of pupae

To establish the major variance in the data prior to further analysis, PCA was applied to the metabolic profiles. Figure 4.4-1-A shows a PCA scores plot of metabolic profile from Control (CONT), Cyp4g16 knock-down (KD16) and Cyp4g17 knock-down (KD17) with PC1 (18.43%) against PC4 (7.40%) with a cumulative variance of 25.83%. In order to explain 95% of the variance in the data, a total of 25 components were required. PC1 and PC4 are PCs showing the highest distance between all three groups compared to other PCs. Each group shows a similar degree of variation within the group as can be seen from the spread of the points in the scores plot. When comparing the metabolic profiles, the PCA scores plot demonstrates a higher degree of similarity between CONT and KD17. This similarity is shown by the majority of the data points being overlapping and spread along the same line in the Figure. In contrast, KD16 samples have less data points overlapping with the rest of the samples and spread along a different axis. Employing a supervised PLS-DA method, metabolites responsible for the metabolic profile differences can be extracted.

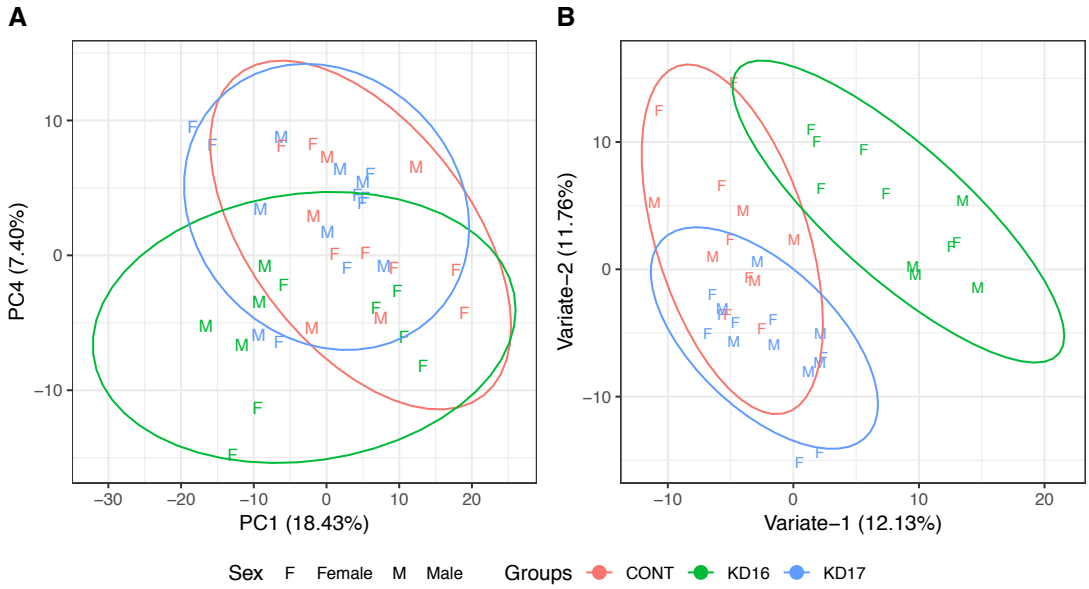


Figure 4.4-1: A) PCA Scores of PC1 (18.43%) against PC4 (7.40%) from Cyp knock-down pupae of *An. gambiae* (n_{KD16}=11, n_{KD17}=15, n_{CONT}=12). A total of 25 PC was used to achieve 95% explained variance. Ellipses represent 95% confidence region. B) Variates one and two of PLS-DA model for knock-down pupae grouped by knock-down status KD16, KD17 and CONT (n_{KD16}=11, n_{KD17}=15, n_{CONT}=12). A higher degree of similarity between KD17 & CONT than to CONT as expected. A three-variate model was determined to be the optimal model complexity *via* cross validation. PLS-DA model average accuracy was calculated as 75.76%. Ellipses represent 95% confidence region.

In order to identify differences in the metabolic profiles, the differences between groups were enhanced using a cross-validated PLS-DA model. Optimal model complexity was found to be a three-variate model with a 75.76% average accuracy (Appendix 11 for further

metrics). Figure 4.4-1-B shows the scores of this model with variate-1 plotted against variate-2 for simplicity. Compared to the PCA plot (Figure 4.4-1-A), a tighter clustering of groups can be observed. A clear separation of KD16 samples from the remaining groups can be seen along a diagonal of variate-1 and variate-2. It was expected for KD16's metabolic profile to be more different than KD17 & CONT. The PLS-DA model achieved a better separation between KD17 and CONT group, although, these groups are not fully decoupled from each other as expected, indicating the higher degree of similarity between the two than to KD16.

4.4.1.2 Key metabolites of pupae

Following a PLS-DA model, metabolites of interest were extracted. In order to accomplish this, VIP scores were calculated for all the bins in the model (Figure 4.4-2). VIP scores for variate-1 and variate-2 were checked for bins scoring higher than the selection threshold of one. A total of 33 identified bins (out of 128) attributed to nine metabolites were selected.

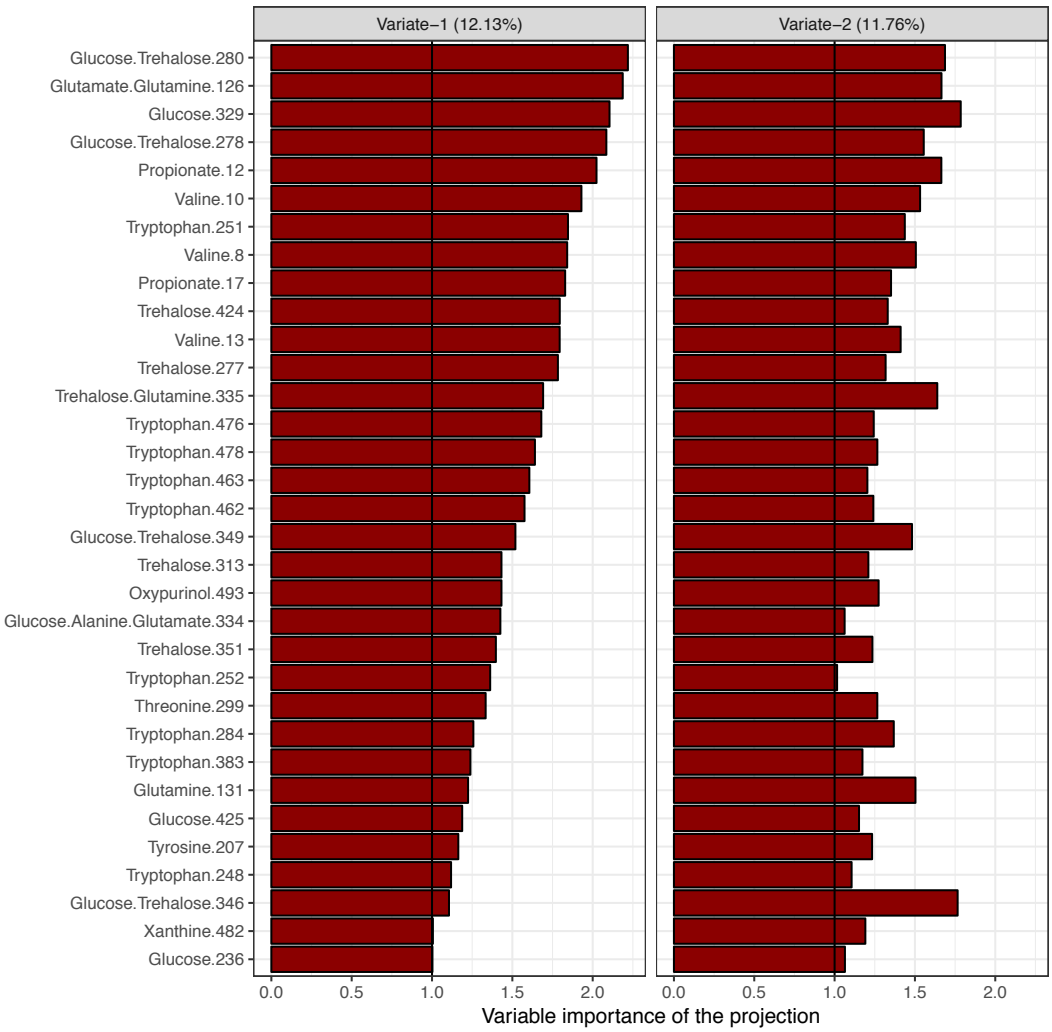


Figure 4.4-2: VIP scores for identified bins calculated from knock-down pupae PLS-DA model discriminating between experimental groups. Bins scoring more than 1 on both variate-1 and variate-2 were selected. Black line represents a VIP score of 1.

CRS was calculated for all identified bins for further shortlisting of metabolite bins. A passing score of 27.52% was calculated using all CRS. VIP selected bins scoring higher than the passing score were considered as a potential representative of its metabolite. Amongst the bins scoring above 27.52%, only non-overlapping (where applicable) bins with the highest score were selected to represent its metabolite. Table 4.4-1 shows the CRS values of VIP selected bins.

Table 4.4-1: CRS for VIP selected bins in PLS-DA model of Cyp knock-down pupae. CRS thresholds were calculated by median(CRS)-SD(CRS). Representative bins were selected according to the highest scoring bin. In the case of an overlapping bin, non-overlapped alternatives were selected where available. Rep: representative bin.

Metabolite	Bin	CRS [%]	CRS > 27.10%	Rep	Metabolite	Bin	CRS [%]	CRS > 27.10%	Rep
Alanine	334*	45.45	✓	334	Tryptophan	476	49.88	✓	476
Glucose	425	57.35	✓	425		478	47.22	✓	
	346*	49.52	✓			462	44.70	✓	
	236	42.72	✓			251	39.85	✓	
	334*	36.05	✓			252	38.82	✓	
	349*	11.81	×			248	37.52	✓	
	280*	6.42	×			463	35.29	✓	
	278*	-3.42	×			383	30.91	✓	
	329	-5.51	×			284	25.34	×	
Glutamate	334*	28.32	✓	334	Tyrosine	207	92.61	✓	207
	126*	0.01	×		Valine	8	68.26	✓	8
Glutamine	335*	26.08	×	-		13	67.81	✓	
	126*	25.91	×			10	67.21	✓	
	131	5.99	×						
Propionate	17	39.48	✓	17					
	12	38.63	✓						
Trehalose	349*	69.92	✓	351					
	351	69.52	✓						
	335*	64.05	✓						
	280*	61.42	✓						
	313	60.94	✓						
	278*	58.56	✓						
	277	55.57	✓						
	424	54.28	✓						
	346*	51.80	✓						

* Denotes overlapping bin.

Applying the CRS scoring threshold on VIP selected bins, eight bins were selected representing eight metabolites. Alanine was only represented with a bin that overlaps with glucose and glutamate. Additionally, the bin representing the methyl group of alanine was not selected, which further indicates that alanine as a metabolite is not a discriminating factor. Hence, alanine was removed from the selected metabolites. Bins representing glutamine did not achieve CRS above 27.10% and so were excluded from further analyses. The selected metabolites (Table 4.4-2) cover three distinct groups of metabolites: amino acids, carboxylic acids and saccharides. In order to verify the discriminatory properties of these metabolites exclusively, metabolomics data was filtered to only include the selected metabolites represented by their respective bins. This selection of metabolites was then analysed by PCA in order to observe major variances represented by these metabolites.

Table 4.4-2: List of selected metabolites from the knock-down pupae PLS-DA discriminating between experimental groups.

Metabolite group	Metabolite	Representative bin	Chemical shift [ppm]	MSI level	KEGG code
Amino acids	Glutamate	334*	3.77	Level 1	C00025
	Tryptophan	476	7.55	Level 1	C00078
	Tyrosine	207	3.05	Level 1	C00082
	Valine	8	1.00	Level 1	C00183
Carboxylic acids	Propionate	17	1.07	Level 2	C00163
Saccharides	Glucose	425	5.24	Level 1	C00031
	Trehalose	351	3.87	Level 1	C01083

* Denotes overlapping bin.

Using the selected metabolites, PCA was performed (Figure 4.4-3-A). PCA scores of PC1 (33.47%) against PC2 (23.11%) account for a total of 56.58% of the explained variance. A total of six components were required in order to explain 95% cumulative variance. The scores plot shows similar results to those obtained in Figure 4.4-1. Some resolution in the group clusters are lost, which indicates there are unidentified metabolites which can discriminate between the experimental groups further. Nevertheless, KD17 and CONT show a higher degree of similarity, with KD16 exhibiting fewer overlapping samples with the spread of the data in the opposite direction on PC2. In order to assess the discriminatory properties of the selected metabolites, a PLS-DA model was built.

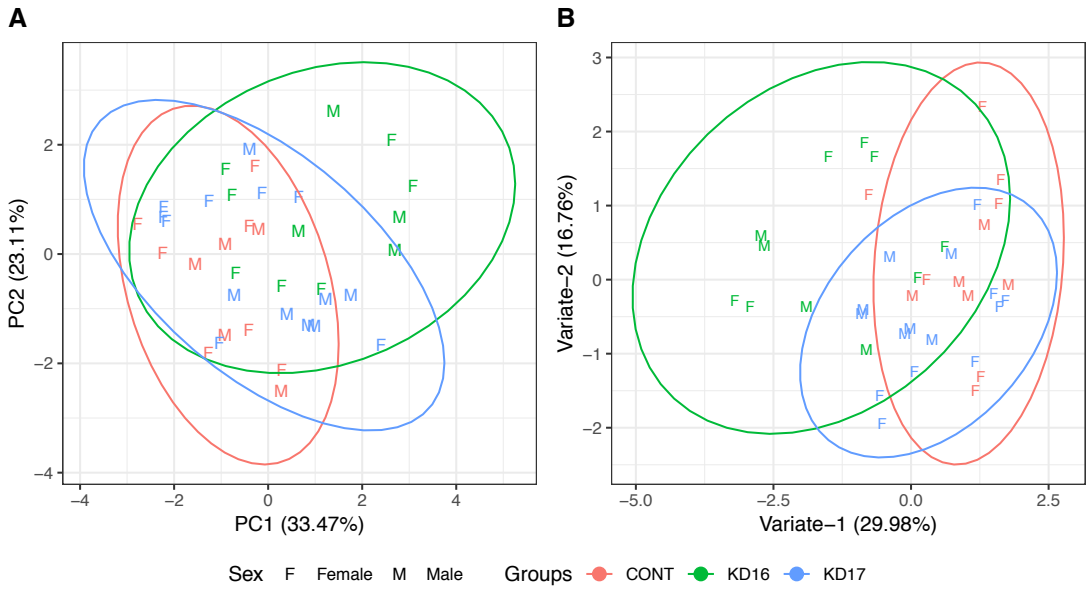


Figure 4.4-3: A) PCA scores plot of selected metabolites for knock-down *An. gambiae* species pupa (n_{KD16}=11, n_{KD17}=15, n_{CONT}=12). A total of 6 components were used to explain 95% variance. Ellipses were drawn using the 95% confidence region. B) PLS-DA scores of knock-down *An. gambiae* pupae (n_{KD16}=11, n_{KD17}=15, n_{CONT}=12). PLS-

DA model complexity was optimised *via* cross-validation, yielding a two-variate model with 69.70% average accuracy. Brackets report on explained variance in its variate. Ellipses represent 95% confidence region.

A cross-validated PLS-DA model was built using the selected metabolites (Figure 4.4-3-B). Optimal model complexity was determined to be a two-variate model *via* cross-validation with 69.79% average accuracy (Appendix 11 for further metrics). PLS-DA scores do not fully separate the three groups, with overlapping clusters similar to the unsupervised PCA (Figure 4.4-3-A). The majority of KD17 and CONT samples were clustered together corresponding to the similarities in their metabolome. Meanwhile, KD16 shows a more unique profile as its clusters are more distinct from the other clusters. In order to probe the selected metabolites further, levels were compared using a BH-adjusted ANOVA followed by a Tukey's HSD pairwise test and shown *via* boxplots.

Metabolite levels were compared *via* univariate ANOVA with BH p-value adjustment (for detailed test statistics see Appendix 17). Metabolites with adjusted p-values less than 0.05 were further analysed by Tukey's HSD test for pairwise comparisons. Figure 4.4-4 shows boxplots of all metabolites selected from the PLS-DA model *via* VIP and CRS scoring. From the selected metabolites only valine, propionate, trehalose and tryptophan were found to be significantly different. More specifically, valine, propionate and trehalose were found to be significantly higher in KD17 than both KD16 and CONT. Finally, tryptophan was significantly higher in KD16 and KD17 compared to CONT. It should be noted that tryptophan presented a sub-population for the KD17 group the sub-population is consisted of four females. Nevertheless, this sub-population is not consistent neither in tryptophan nor other selected metabolites. It is clear that even with the four females clustering, the rest of the data is a mix of males and females.

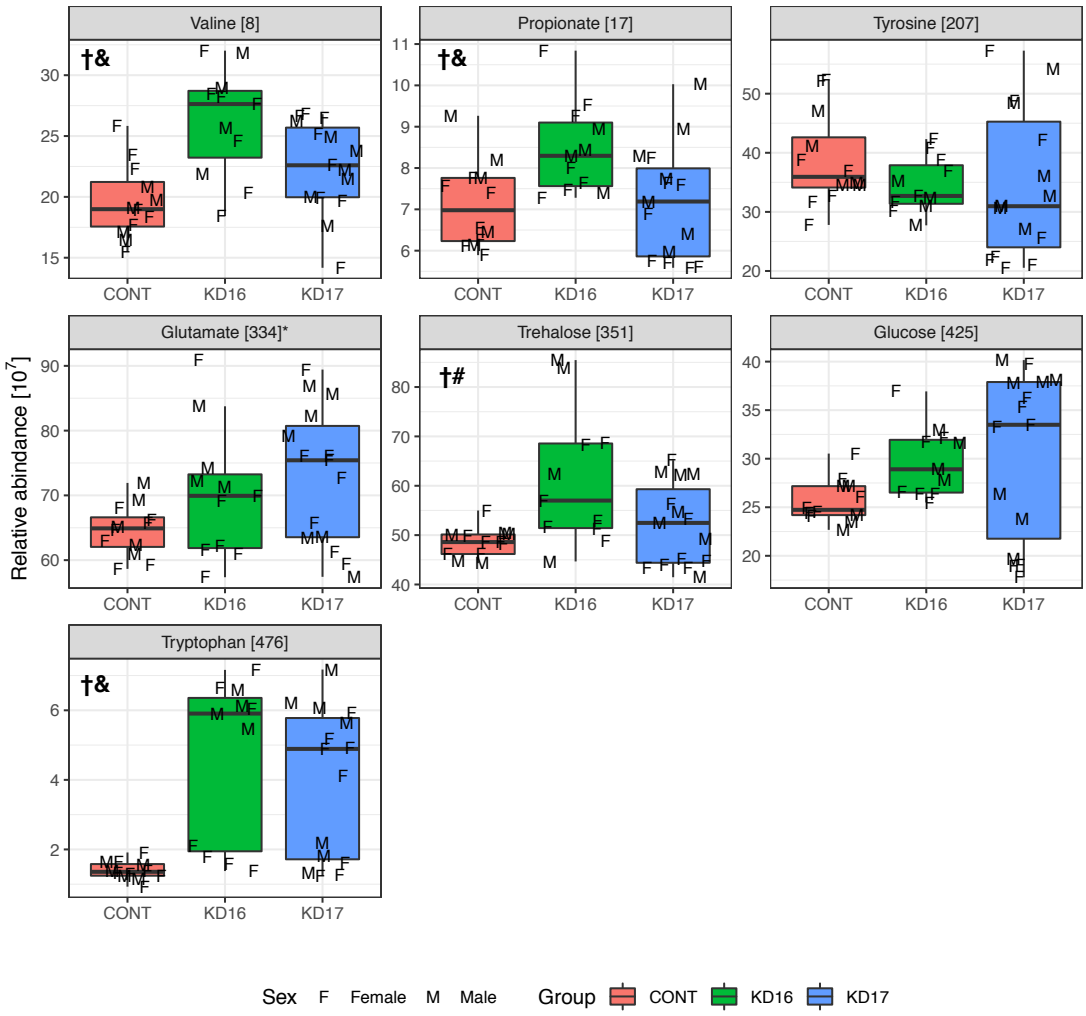


Figure 4.4-4: Boxplots of selected metabolites, square brackets indicates the representative bin used for the metabolite. Significant difference between pairs are denoted by the following symbols: †, KD16 and CONT; &, KD17 and CONT; #, KD16 and KD17.

4.4.2 Analysis of adults metabolic profile

4.4.2.1 Statistical analysis of adults

The analytical approach taken in section 4.4.1 was applied to the adult data as well. The unsupervised analysis *via* PCA was performed in order to determine the existence of underlying structure corresponding to the knock-downs in the data. The PCA scores (Figure 4.4-5-A) plot shows PC1 (30.11%) against PC2 (14.02%) which demonstrates the most variance between groups accounting for a cumulative variance of 44.13%. PCA transformation of the adult data achieved 95% explained variance with 48 components. The PCA plot shows a higher degree of metabolic profile similarity between CONT and KD16 revealed by the overlapping of data points as well as the similar spread along PC1. KD17 are more tightly clustered along PC1 presenting a wider spread along PC2 as well, KD17 samples

exhibit less overlap than KD16 and CONT samples. Similar to the pupae data sub-population can be seen, interestingly to a lesser degree.

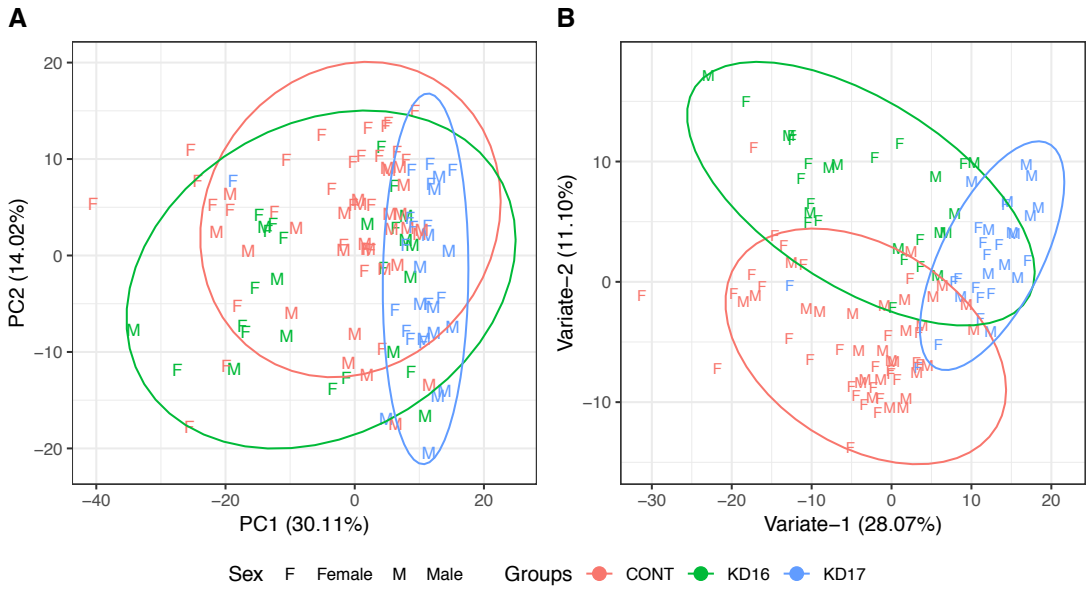


Figure 4.4-5: A) PCA scores plot for the knock-down *An. gambiae* adults (n_{CONT}=56, n_{KD16}=27, n_{KD17}=29). A total of 48 components were required to account for 95% of the explained variance in the data. Ellipses were drawn using the 95% confidence interval of each group. B) PLS-DA model for knock-down *An. gambiae* adults (n_{CONT}=56, n_{KD16}=27, n_{KD17}=29). Model was built with 2 variates, chosen by cross-validation with 82.35% average accuracy. Brackets report its variates explained variance. Ellipses were drawn using the 95% confidence interval of each group.

To explore the variation in metabolic profiles further, a supervised PLS-DA model was built with cross-validation. Figure 4.4-5-B shows the PLS-DA model discriminating between the KD16, KD17 and CONT groups. A two-variate model was selected to be the optimal model through cross-validation with 82.35% average accuracy (Appendix 11 for further metrics). PLS-DA scores plot (Figure 4.4-5-B) showed closer clustering of KD17 samples compared to KD16 and CONT groups, suggesting variation to a lesser degree out of all the experimental groups. This was expected due to the high activity of Cyp4g17 in adults. In terms of metabolic profile similarities, the clusters are tighter compared to what was observed in the PCA plot (Figure 4.4-5-A). Even after the PLS-DA model a sub-population can still be observed, suggesting the cause of this variation is strong. To probe the metabolites most influential on this discrimination, VIP scores were calculated.

4.4.2.2 Key metabolites of adults

Metabolite selection is critical for understanding the underlying processes of the knock-downs. In order to achieve this, most influential spectral features representing the metabolites were identified from the PLS-DA model. Bins representing spectral features were

ranked *via* VIP scores. A threshold for selection was applied, where only bins scoring higher than 1 were selected. Since this model's discriminative features were expressed on variate-1 and variate-2, bins were required to score higher than 1 on both variates. Figure 4.4-6 shows the bins satisfying the criteria mentioned above. Only 100 bins out of 496 (20.16%) scored higher than the threshold. Out of the 100, only 11 (11.00%) were identified representing seven unique metabolites. Due to the nature and complexity of NMR, prior to metabolite selection, each bin needs to be ensured of it its representative quality. This was performed *via* CRS.

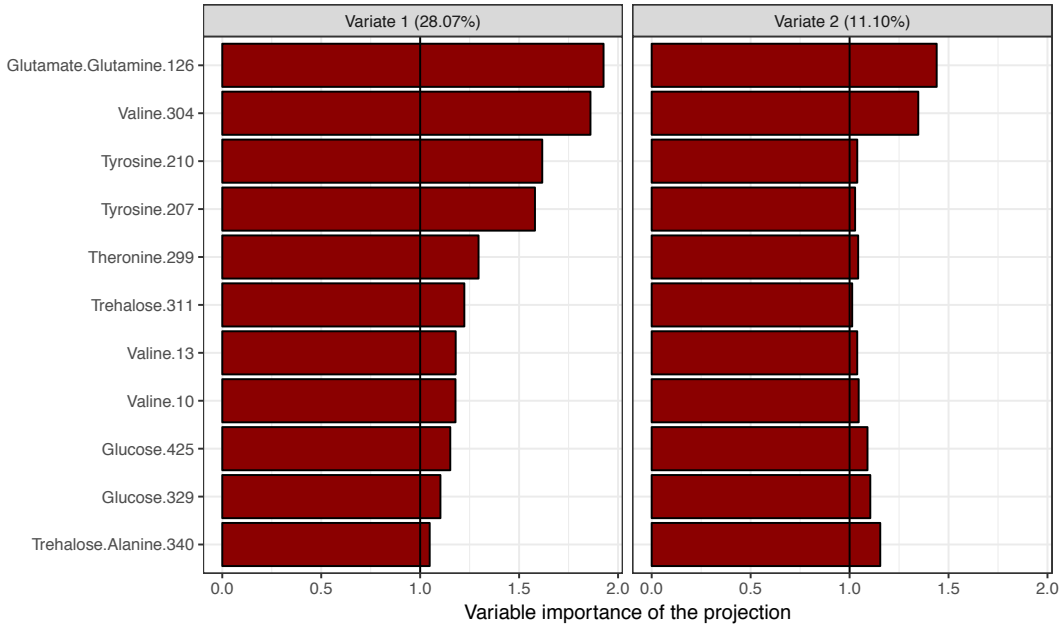


Figure 4.4-6: VIP scores for knock-down *An. gambiae* adults with an applied threshold of 1. Bins scoring more than 1 on variate-1 and variate-2 were selected for further verification. A total of 11 (identified only) out of 496 (2.22%) bins were selected. Black line represents a VIP score of 1.

CRS provides a reliability measure using the correlation values of the NMR spectra. A CRS passing score of 29.14% was calculated using all CRS calculated from the identified bins. The adult dataset yielded a shorter list of 11 bins compared to the pupa VIP list of 33 bins. Using CRS (Table 4.4-3), 10 bins selected from VIP scoring bins were assessed for their metabolite representation quality. The highest scoring, non-overlapping (where available) bins above the passing score were selected.

Table 4.4-3: CRS scores of VIP selected bins. Threshold of 29.14% was calculated by median(CRS)-SD(CRS). Rep: representative bin.

Metabolite	Bin	CRS [%]	CRS > 29.14%	Rep	Metabolite	Bin	CRS [%]	CRS > 29.14%	Rep
Alanine	340*	67.47	✓	340	Tyrosine	207	59.30	✓	207
Glucose	329	90.23	✓	329		210	59.05	✓	
	425	88.96	✓		Valine	10	46.98	✓	10
Glutamate	126*	-6.16	×	-		13	46.54	✓	
Glutamine	126*	7.94	×	-		304	12.94	×	
Trehalose	340*	61.45	✓	311					
	311	52.89	✓						

* Denotes overlapping bin.

Following the CRS, the metabolites glutamate and glutamine failed to score above the passing score of 29.14% and so these metabolites were excluded from the list of selected metabolites (Table 4.4-4). PCA was performed on the selected representative bins in order to determine the influence of these metabolites in relation to the profiles of the experimental groups.

Table 4.4-4: Shortlist of selected metabolite from the PLS-DA model by VIP scores and CRS filtering.

Metabolite group	Metabolite	Representative bin	Chemical shift [ppm]	MSI level	KEGG code
Amino acids	Alanine	340*	3.79	Level 1	C00041
	Tyrosine	207	3.05	Level 1	C00082
	Valine	10	1.04	Level 1	C00183
Saccharides	Glucose	329	3.74	Level 1	C00031
	Trehalose	311	3.65	Level 1	C01083

* Denotes overlapping bin.

Using the dataset including only the metabolites represented by their selected bins, PCA was performed. Figure 4.4-7-A shows the PCA scores plot of PC1 (60.29%) against PC2 (18.87%) explaining a cumulative variance of 79.16%. The scores plot shows a tighter clustering of KD17 samples compared to KD16 and CONT samples. KD17 samples do not show high overlap with KD16 and CONT samples suggesting a differing metabolic profile; this is consistent with the previous PCA performed with the entire data set (Figure 4.4-5-A). KD16 and CONT samples possess higher numbers of overlapping samples and similar variation spread which are also consistent with the previously observed PCA plot (Figure 4.4-5-A). Interestingly, KD16 and CONT share the same variance spread along PC1 but, their spread on PC2 are opposite suggesting minor differences between the two metabolic profiles. Interestingly the sub-population in KD16 is retained. Given the high variation observed in Figure 4.4-5-A, it is not surprising, what is not expected is the model could not minimize the within group

variation as expected. Further discriminatory properties of these metabolites were demonstrated *via* supervised PLS-DA.

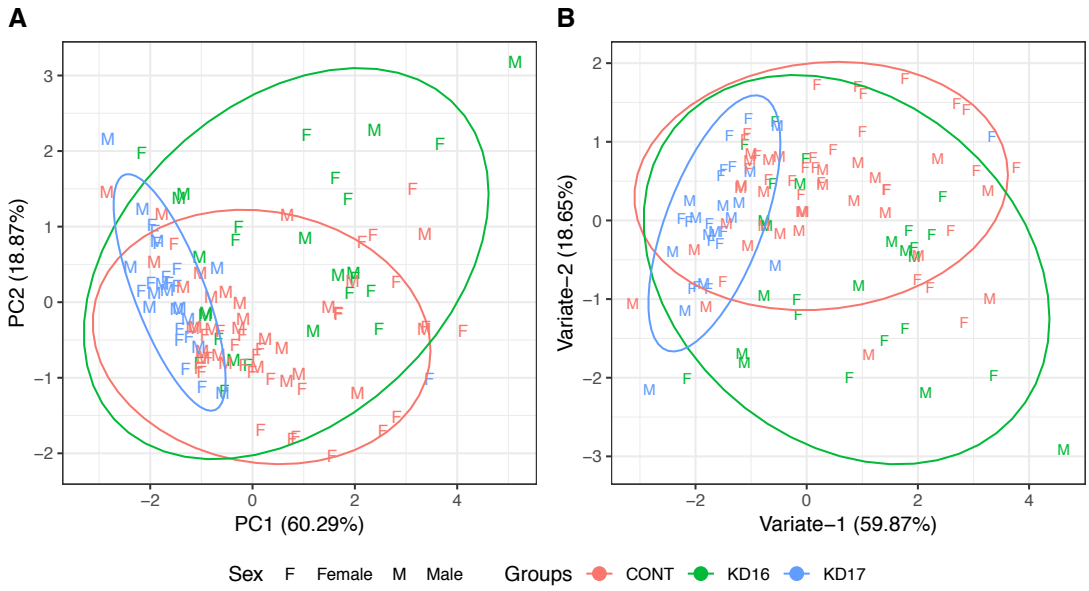


Figure 4.4-7: A) PCA scores plot of selected metabolites of knock-down *An gambiae* adult (n_{CONT}=56, n_{KD16}=27, n_{KD17}=29). Only selected bins were used as for the PCA. A total of four components were used to explain 95% variance. Ellipses represent the 95% confidence region. B) PLS-DA model for knock-down *An. gambiae* adults (n_{CONT}=56, n_{KD16}=27, n_{KD17}=29). Model was built with two variates, chosen by cross-validation. Model average accuracy was calculated as 76.74%. Variate's explained variance are reported in the brackets. Ellipses represent 95% confidence region.

Using the selected metabolite bins a cross-validated PLS-DA model (Figure 4.4-7-A) was built. The result of optimistaion is a two-variate model with 76.74% aveage accuracy (Appendix 11 for further metrics). PLS-DA scores plot exhibits a similar clustering pattern to that observed in the previous PCA and PLS-DA plots (Figure 4.4-5-A, Figure 4.4-5-B, Figure 4.4-7-A). As observed previously, KD17 samples clustered tighter and were positioned in a more separated manner compared to KD16 and CONT groups. In order to reveal molecular level information on exhibited differences, metabolite levels were compared using BH adjusted ANOVA followed by Tukey's HSD test for pairwise comparison (for detailed test statistics see Appendix 17).

From the selected metabolites, CHC biosynthesis precursor valine was found significantly higher in both KD16 and KD17 compared to CONT. Similarly, Tyrosine was also significantly higher in KD17 compared to both KD16 and CONT. The amino acid alanine was found significantly lower compared to both KD16 and CONT. The remaining selected metabolites were the saccharides glucose and trehalose, and both were significantly lower in KD17 compared to KD16 and CONT groups. The boxplots shows the source of the variation

observed in the PCA (Figure 4.4-5-A) and PLS-DA plots (Figure 4.4-5-B). Glucose and alanine shows a very strong formations of sub-population not attributed to sex. When compared to the spectra (Appendix 16) a distinctly high signal intensities can be observed in the region where glucose signals arise for the KD16 group. Similarly the selected alanine signal is an overlapping one although, it scored a passing CRS score it might be suffering from the variation caused by the overlapping signal or the signals around it.

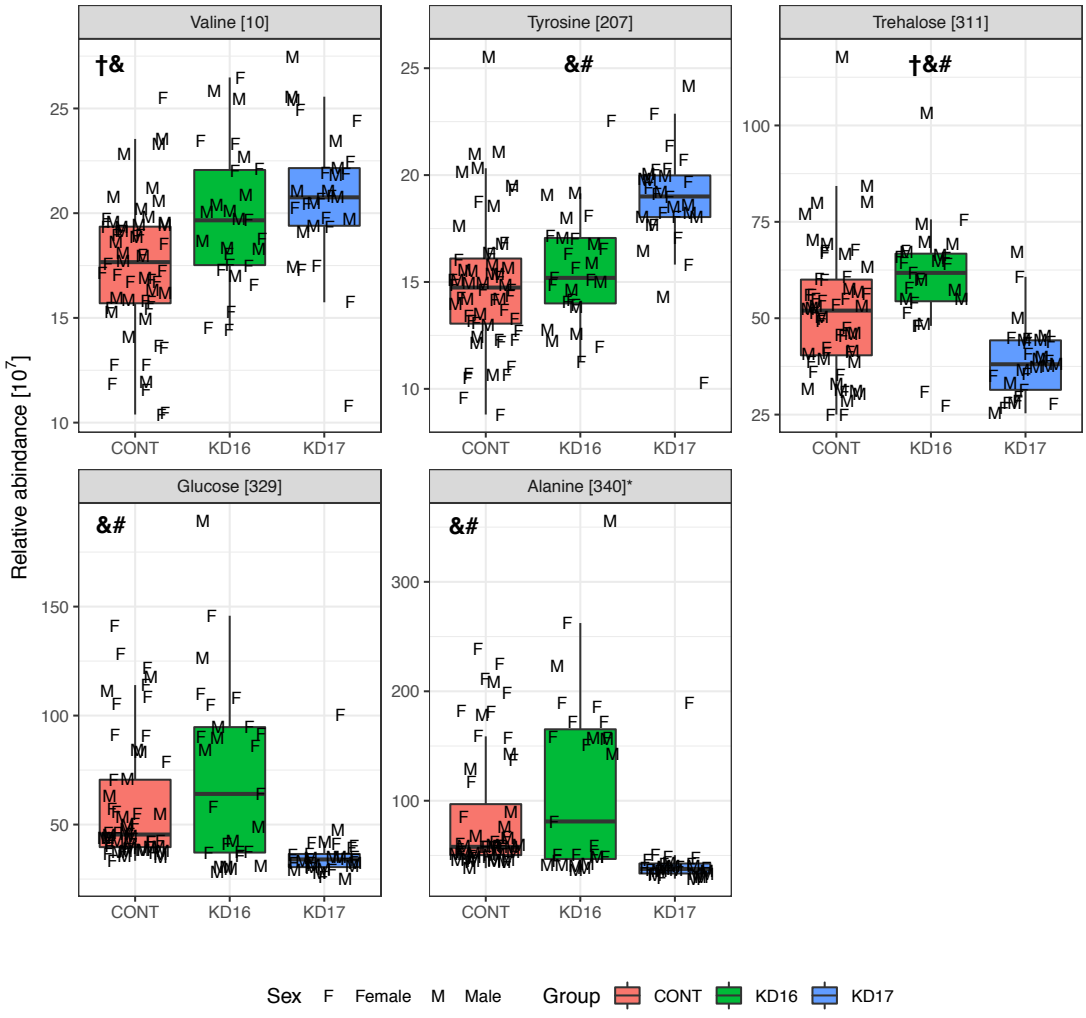


Figure 4.4-8: Boxplots of metabolite selected by VIP and CRS scores (n_{CONT}=56, n_{KD16}=27, n_{KD17}=29). CONT, control; KD16, knock-down of Cyp4g16; KD17, knock-down of Cyp4g17. Significant difference between pairs are denoted by the following symbols: †, KD16 and CONT; &, KD17 and CONT; #, KD16 and KD17.

4.4.3 Metabolite set enrichment analysis

Selected metabolites for both stages in knock-down *An. gambiae* were subjected to metabolite set enrichment analysis (MSEA) to extract further metabolic pathway level information from both stages. Table 4.4-5 summarises the metabolites selected for both pupa and adult. These metabolites belong to three classes; amino acids, carboxylic acids and saccharides.

Table 4.4-5: Summary of selected metabolites of knock-down *An. gambiae* in pupae and adults. Levels are shown qualitatively for simplicity, although metabolites were compared quantitatively *via* Tukey's HSD. ↑, significantly higher; ↓, significantly lower, NS (arrow) denotes non-significant change with mean abundance level, and square brackets represent BH-adjusted p-values.

Metabolite		**Levels compared to Control**				
class	**Pupa**	**KD16**	**KD17**	**KD16**	**KD17**	**Adult**
Amino acids	Alanine	-	-	NS (↑) [6.31×10^{-2}]	↓ [6.31×10^{-3}]	Alanine
	Glutamate	NS (↑) [2.85×10^{-1}]	NS (↑) [5.23×10^{-2}]			
	Tryptophan	↑ [5.63×10^{-4}]	↑ [2.85×10^{-3}]			
	Tyrosine	NS (↓) [5.91×10^{-1}]	NS (↓) [5.14×10^{-1}]	NS (↑) [5.91×10^{-1}]	↑ [2.44×10^{-7}]	Tyrosine
	Valine	↑ [5.35×10^{-4}]	NS (↑) [1.29×10^{-1}]	↑ [5.71×10^{-3}]	↑ [3.66×10^{-5}]	Valine
Carboxylic acids	Propionate	↑ [2.66×10^{-2}]	NS (↑) [9.91×10^{-1}]			
Saccharides	Glucose	NS (↑) [2.28×10^{-1}]	NS (↑) [9.24×10^{-2}]	NS (↑) [1.13×10^{-1}]	↓ [3.04×10^{-3}]	Glucose
	Trehalose	↑ [5.69×10^{-3}]	NS (↑) [5.90×10^{-1}]	↑ [3.48×10^{-2}]	↓ [6.75×10^{-4}]	Trehalose

MSEA was performed on selected metabolites using a database curated from the KEGG pathways (*An. gambiae* PEST strain [KEGG organism code: aga]). MSEA was performed using Fisher's exact test with EASE correction and BH p-value adjustment for multiple testing (

Table 4.4-6). Pathways significantly over-represented *via* MSEA (Figure 4.4-9) are biosynthesis of amino acids, aminoacyl-tRNA biosynthesis and ABC transporters.

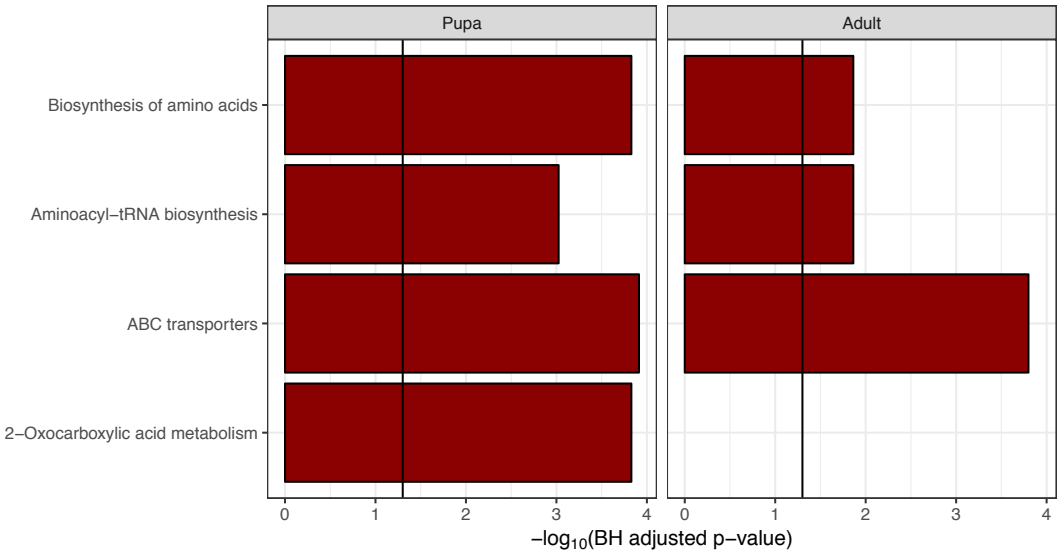


Figure 4.4-9: MSEA showing over-represented pathways from the selected metabolites influencing the discrimination between KD16, KD17 and CONT groups. Black line represents p-value of 0.05.

Table 4.4-6: Match table for identified MSEA pathways for three-way comparison of KD *An. gambiae*.

Pathway	Stage	Raw p-value	BH adjusted p-value	Hit/total (%)	Matches
2-Oxocarboxylic acid metabolism	Pupa	0.000014	0.000148	4/134 (2.99%)	Glutamate, Tryptophan, Tyrosine, Valine
ABC transporters	Pupa	0.000003	0.000122	4/182 (2.2%)	Glutamate, Glucose, Valine, Trehalose
	Adult	0.000006	0.000159	4/182 (2.2%)	Glucose, Alanine, Valine, Trehalose
Aminoacyl-tRNA biosynthesis	Pupa	0.000108	0.000945	4/52 (7.69%)	Glutamate, Tryptophan, Tyrosine, Valine
	Adult	0.001643	0.013692	3/52 (5.77%)	Alanine, Tyrosine, Valine
Biosynthesis of amino acids	Pupa	0.000010	0.000148	4/128 (3.13%)	Glutamate, Tryptophan, Tyrosine, Valine
	Adult	0.001423	0.013692	3/128 (2.34%)	Alanine, Tyrosine, Valine

4.5 Analysis of specificity of Cyp4g16 and Cyp4g17 knock-down by metabolic profiling

Following the metabolic profile comparisons of KD16, KD17 and CONT groups, it was found that both knock-down groups differ from the CONT particularly at different life stages.

A temporal difference between KD16 and KD17 was expected as it was observed during breeding and as reported by Lynd et al [188]. In section 4.4.1 it was observed that KD16 pupae's metabolic profile is different in comparison to pupae of KD17 and CONT which was attributed to the early expression of Cyp4g16 in pupal development. Similarly, a difference was observed in KD17 adults, which was attributed to the Cyp4g17 expression being predominant in late pupa/adult stages. This section focuses on exploring the differences between knock-downs of Cyp4g16 and Cyp4g17. A pairwise comparison of KD16 and KD17 aims to understand how Cyp4g16 and Cyp4g17 affect the pupal development differently. A comparison of KD16 and KD17 will be made both for pupa and adult stages. Resulting metabolites will be used in a MSEA and the results will be compared.

4.5.1 Analysis of specificity in pupae

4.5.1.1 Statistical analysis

In order to observe major variances between KD16 and KD17, PCA (Figure 4.5-1-A) was performed. PC1 (19.86%) and PC4 (8.75%) show the differences between KD16 and KD17 metabolic profiles clearest compared to other PCs. A total of 18 components were required to explain the 95% variance in the data, with PC1 and PC4 accounting for a cumulative variance of 28.61%. The difference between KD16 and KD17 metabolic profiles were most distinct on PC4 whereas, on PC1 both groups followed a similar spread.

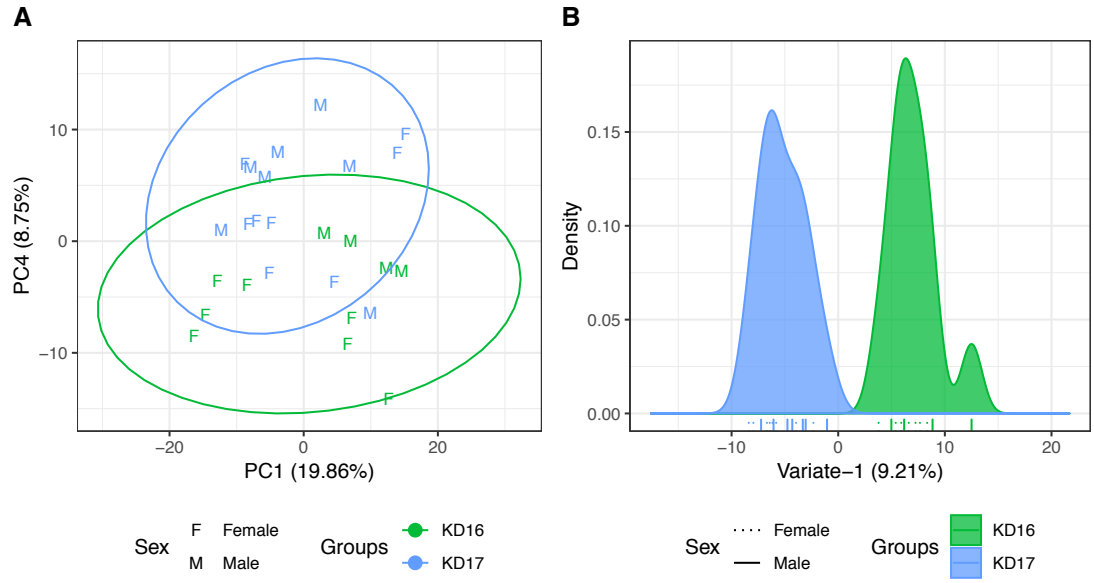


Figure 4.5-1: A) PCA scores plot for knock-down pupae KD16 vs KD17, PC1 against PC4 (n_{KD16}=11, n_{KD17}=15). A total of 18 components were required to explain 95% of the variance in the data. Ellipses represent 95% confidence region of each group. B) PLS-DA density plot of Variate-1, discriminating between KD16 and KD17 pupae (n_{KD16}=11, n_{KD17}=15). Single-variate (9.21% explained variance) model complexity (75.00% accuracy) was optimised with cross-validation. Each tick represents a sample from KD16 (green) or KD17 (blue) and tick type represents sex straight for male and dashed for female.

The PCA scores plot suggest KD16 and KD17 metabolic profiles are similar and their differences are represented by a small proportion of the variance in the data (i.e. PC4 8.75%). Using a PLS-DA model, the variation within groups can be suppressed while accentuating the variation between groups, hence highlighting subtle differences between KD16 and KD17. In order to build the PLS-DA model, optimisation was performed *via* cross-validation. A single variate model was identified as the optimal model complexity with 75.00% accuracy (Appendix 11 for further metrics). Figure 4.5-1-B shows the density plot of the PLS-DA model. The ticks shown under the density plots show a clear separation of the KD16 and KD17 groups. In order to probe the metabolites most influential on the discrimination between KD16 and KD17, VIP scores were calculated.

4.5.1.2 Key metabolites of the comparison

VIP scores were calculated for all bins from the PLS-DA model. A passing threshold of one was applied to all scores. Out of 496 bins, only 170 (34.27%) bins scored higher than the threshold. From the 170 bins, only 41 (24.12%) were identified. Figure 4.5-2 shows the 41 identified bins scoring higher than the passing threshold which represent 11 unique metabolites. To achieve robust metabolite selection, CRS were calculated for the identified bins and then used to perform metabolite selection from the bins with VIP scores higher than one.

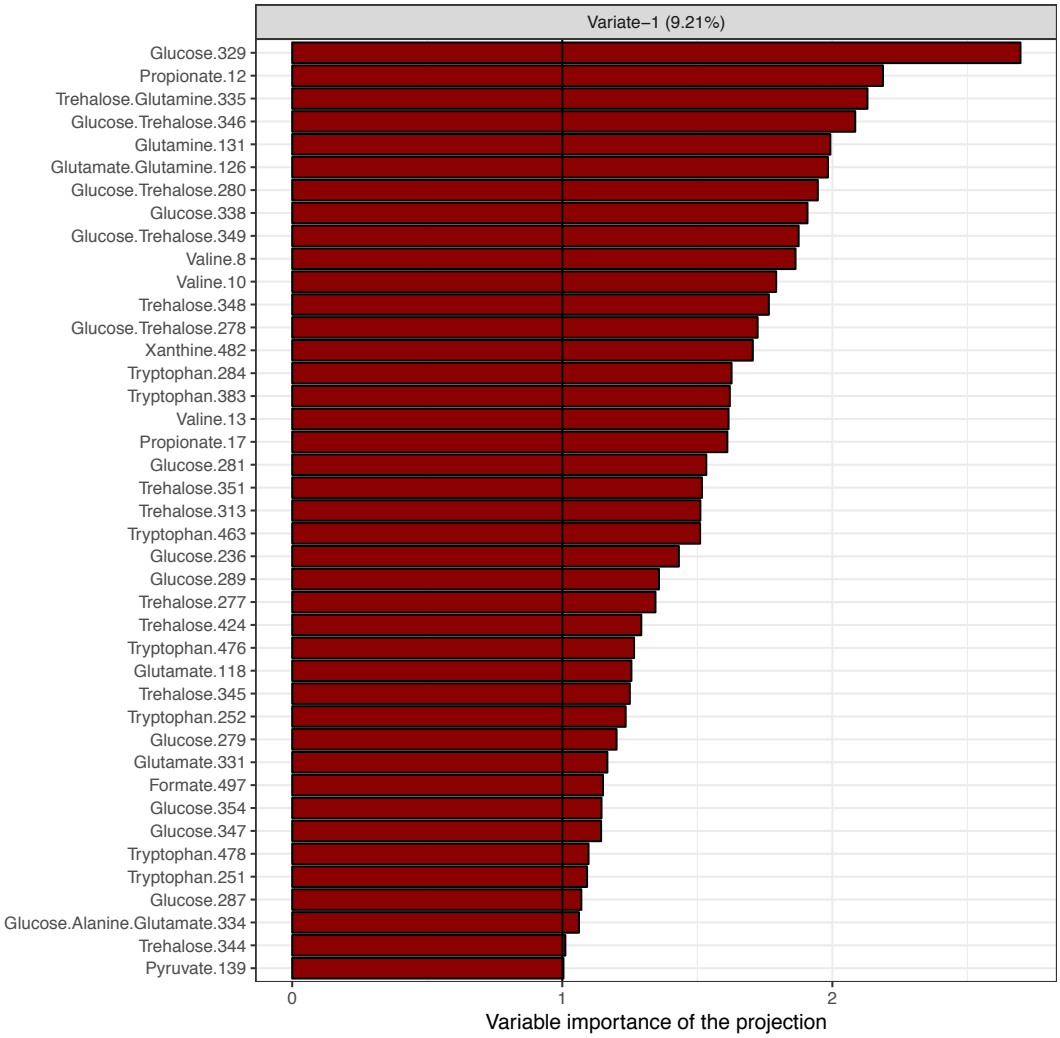


Figure 4.5-2:VIP scores of identified bins scoring higher than one from the PLS-DA model build for discriminating between KD16 and KD17 pupae. The list contains 41 bins representing 11 unique metabolites. The total number of bins in the model was 496, of which 170 scored higher than one. Black line represents VIP score of 1.

CRS was calculated from the correlations of all the identified bins (114 bins in total), and a passing score of 25.77% was calculated from all the CRS (Table 4.5-1). Amongst the bins passing the threshold, only the highest, non-overlapping (where applicable) were selected.

Table 4.5-1:CRS of selected bins *via* VIP for KD16 vs KD17 pupae. A minimum of 25.77% (median(CRS)-sd(CRS)) was required for a bin to be considered as representative. Unique (non-overlapping) peaks were selected where available.

Metabolite	Bin	CRS [%]	CRS > 25.77%	Rep	Metabolite	Bin	CRS [%]	CRS > 25.77%	Rep
Alanine	334*	41.75	✓	334	Trehalose	351	71.12	✓	351
Formate	497	Singlet	NA	497		349*	71.01	✓	
Glucose	354	66.41	✓	354		344	68.05	✓	
	279	65.49	✓			345	67.79	✓	
	281	63.69	✓			335*	63.78	✓	
	287	63.20	✓			280*	63.29	✓	
	289	62.42	✓			278*	61.90	✓	
	347	52.46	✓			313	59.92	✓	
	346*	45.27	✓			348	59.81	✓	
	236	42.51	✓			277	59.47	✓	
	334*	38.35	✓			424	59.27	✓	
	338	31.21	✓			346*	52.14	✓	
	280*	3.09	×		Tryptophan	476	56.17	✓	476
	349*	3.08	×			478	54.52	✓	
	329	-5.08	×			251	49.51	✓	
	278*	-5.79	×			252	42.36	✓	
Glutamate	331	48.15	✓	331		463	40.72	✓	
	118	31.82	✓			383	38.15	✓	
	334*	28.32	✓			284	30.45	✓	
	126*	0.07	×		Valine	13	65.77	✓	13
Glutamine	126*	28.58	✓	126		10	65.10	✓	
	335*	27.14	✓			8	64.80	✓	
	131	4.71	×		Xanthine	482	Singlet	NA	482
Propionate	17	45.62	✓	17					
	12	39.94	✓						
Pyruvate	139	Singlet	NA	139					

*: Overlapping bin

Following CRS, 11 bins were selected as representatives for individual metabolites with alanine and glutamine representatives being the only overlapping bins. Within the overlapping bins, each were represented in the list only once. Selected metabolites covered four classes: amino acids (alanine, glutamate, glutamine, tryptophan and valine), carboxylic acids (formate, propionate and pyruvate), purines (xanthine), and saccharides (glucose and trehalose).

Table 4.5-2: Bins selected to represent metabolites influencing the difference between KD16 and KD17 pupae in the PLS-DA model.

Class	Metabolite	Bin	Chemical shift [ppm]	KEGG code
Amino acids	Alanine	334*	3.77	C00041
	Glutamate	331	3.76	C00025
	Glutamine	126*	2.14	C00064
	Tryptophan	476	7.55	C00078
	Valine	13	1.05	C00183
Carboxylic acids	Formate	497	8.46	C00058
	Propionate	17	1.07	C00163
	Pyruvate	139	2.37	C00022
Purines	Xanthine	482	7.89	C00385
Saccharides	Glucose	354	3.89	C00031
	Trehalose	351	3.87	C01083

Prior to pathway analysis and biological contextualisation, raw data was filtered to include only the representative bins. PCA was performed on the filtered data to observe the major variances and their correspondence to KD16 and KD17 profiles. PCA scores (Figure 4.5-3-A) of PC1 (33.77%) against PC3 (17.09%) show the metabolic profile differences between KD16 and KD17 best. A total of 8 components were required to explain 95% of the variance between KD16 and KD17 with, PC1 and PC3 accounting for 50.86% of this variance. In order to elucidate the influence of selected metabolites over the discrimination of the groups individually, unsupervised PLS-DA was applied.

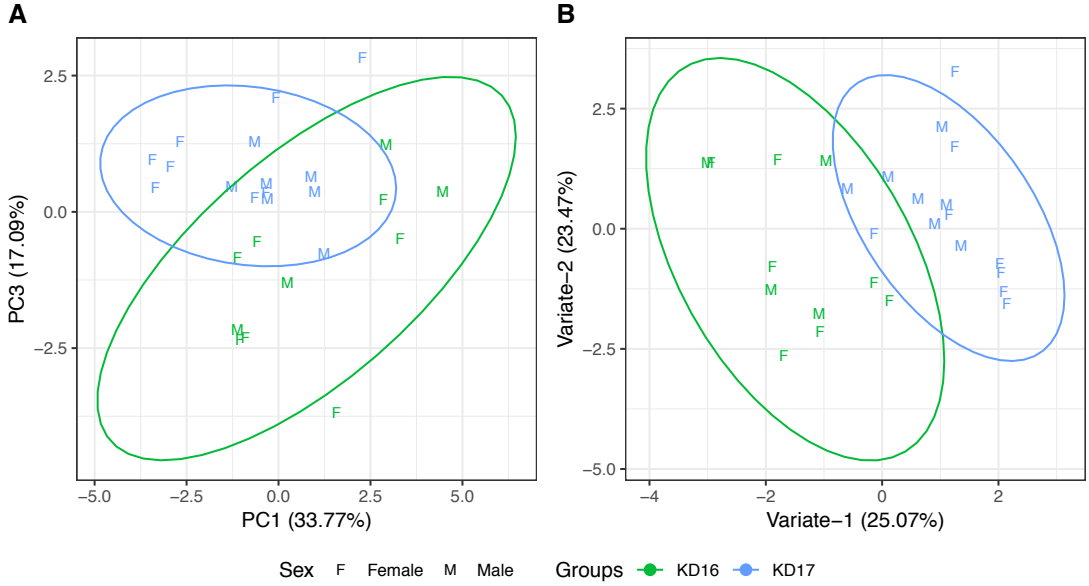


Figure 4.5-3: A) PCA scores plot selected PCA of knock-down *An. gambiae* pupae (n_{KD16}=11, n_{KD17}=15). Only selected bins were used as for the PCA. A total of eight components to explain 95% of the variance. Ellipses represent 95% confidence region for each group. B) PLS-DA model built with only representative bins discriminating between KD16 and KD17 pupae (n_{KD16}=11, n_{KD17}=15). Model complexity was determined to be two-variates *via* cross validation with 75.00% accuracy. Explained variance for each variate are reported in the brackets. Ellipses represent 95% confidence region.

A PLS-DA model of the filtered dataset (Figure 4.5-3-B) was built with cross-validation, optimised to a two-variate model with 75.00% accuracy (Appendix 11 for further metrics). PLS-DA scores plot shows separation for the majority of the data points between KD16 and KD17 along a diagonal of variate-1 and variate-2. In order to probe the metabolite levels responsible for the differences between KD16 and KD17, metabolite levels were compared using BH adjusted t-tests.

The levels of selected metabolites between KD16 and KD17 were visualised using boxplots (Figure 4.5-4). The boxplots of KD16 and KD17 shows clearer picture on forming sub-populations. Although it is evident that these sub-populations are on certain metabolites and not the others. It is consistently more pronounced in KD16 samples than KD17. In order to compare the levels of selected metabolites between KD16 and KD17, a BH p-adjusted t-test was applied (for detailed test statistics see Appendix 17). Differences in metabolites valine, propionate, glutamine, glutamate, trehalose and xanthine were observed, although none were significant after BH-adjustment (Appendix 17).

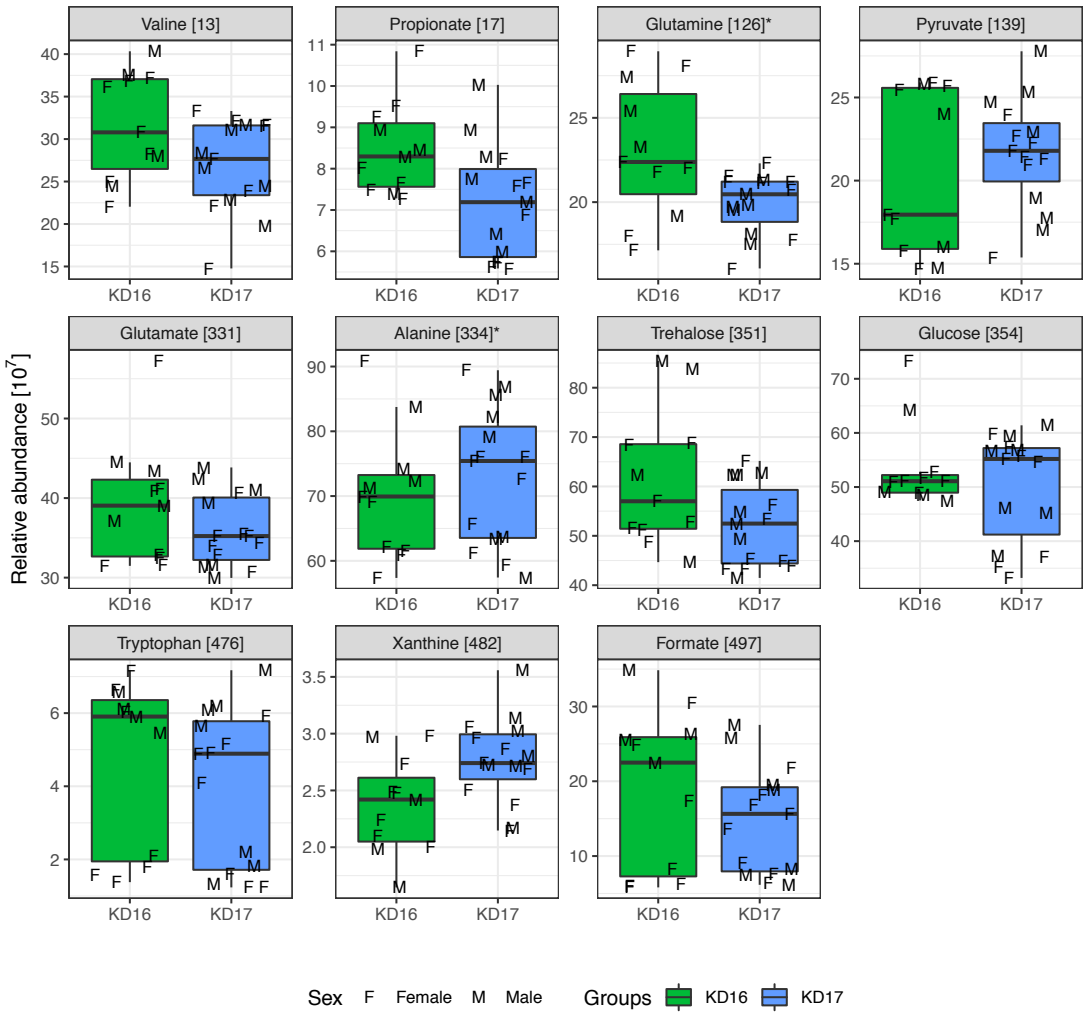


Figure 4.5-4: Boxplots from pupa samples showing the levels of selected metabolites from the PLS-DA model (n_{KD16}=11, n_{KD17}=15). Metabolite levels were compared using a BH adjusted t-test. * by the metabolite name denotes overlapping bin.

4.5.2 Analysis of specificity in adults

4.5.2.1 Statistical analysis

The adult dataset analysis was performed with the same approach as for the pupal analysis. PCA was performed on the adult data to observe the major variances in the data and how they correspond. Figure 4.5-5-A shows PC1 (35.33%) and PC4 (6.01%) accounting for a cumulative variance of 41.34%. To explain the 95% variance in the data, a total of 30 components were required. PC1 and PC4 were found to be the components representing greatest metabolic difference between KD16 and KD17. The KD17 group exhibits a close clustering both on PC1 and PC4. In comparison, KD16 is more variant in both PC1 and PC4. Due to the comparatively variance of the KD16 samples. Although KD17 demonstrates a tightly clustered cohort, two samples were found far away from their clusters. These two samples, one being separated along PC1 and the other along PC4, are mostly likely to be

extreme cases of their group. According to the available information on the samples, these differences could not be attributed to sex, batch, or controlled environmental factors such as light cycle, humidity and temperature.

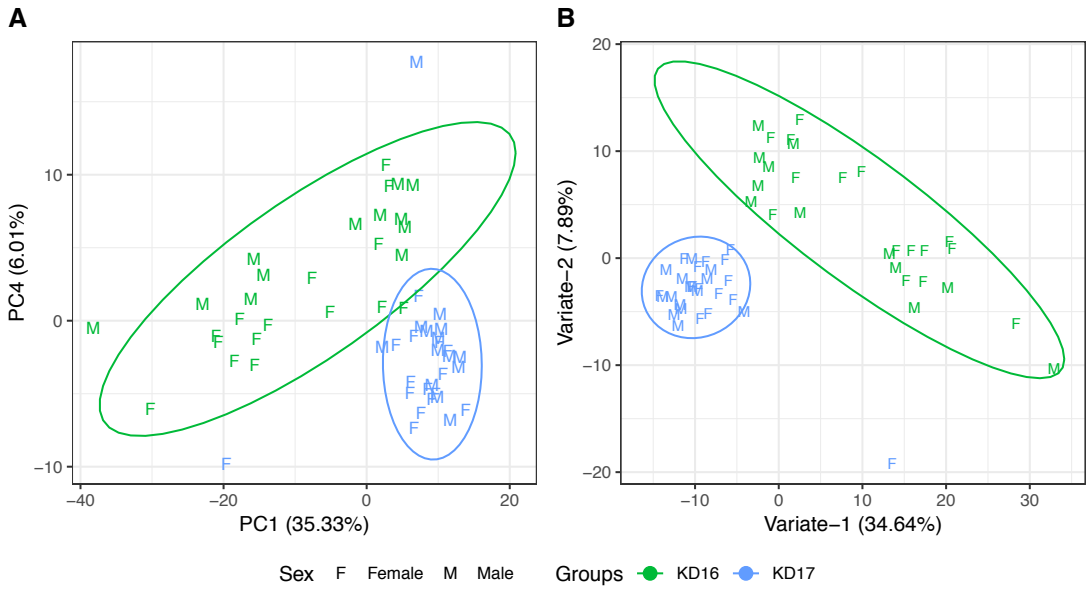


Figure 4.5-5: A) PCA scores plot for knock-down adults KD16 vs KD17. Plotted as PC4 (6.01%) against PC1 (35.33%) showing the metabolic difference between KD16 and KD17 accounting for a cumulative variance of 41.34%. A total of 30 components were needed to explain 95% variance in the data. Ellipses represent the 95% confidence region of each group. B) PLS-DA model of adult knock-down *An. gambiae* (n_{KD16}=27, n_{KD17}=29). Cross-validation was used to determine the two-variate complexity of the model with 100% accuracy. Variate-1 (34.64%) and variate-2 (7.89%) cumulative variance is 42.53%. Ellipses represent 95% confidence region.

As shown by the PCA plot, metabolic differences between KD16 and KD17 could be attributed to the 41.34% of the variance in the data. Using a cross validated PLS-DA model, metabolic differences between KD16 and KD17 can be accentuated further allowing the selection of the most influential spectral features discriminating the two groups. A cross-validated PLS-DA model was built (Figure 4.5-5-B) using cross-validation resulting in a two-variate model complexity with 100% accuracy (Appendix 11 for further metrics). The model shows a clear separation between KD16 and KD17 along a diagonal of variate-1 and variate-2 of the scores plot. The KD17 group exhibited a tighter clustering compared to KD16, as was observed in the previous PCA scores plot (Figure 4.5-5-A). To obtain metabolite level information from the model, VIP scores were calculated.

4.5.2.2 Key metabolites of comparison

The importance of each bin projection from the PLS-DA model was calculated to select the most influential spectral features in the model. Only identified bins scoring higher than 1 on both variate-1 and variate-2 were selected. VIP scores were calculated for all 496 bins. A total

of 189 bins scored higher than the VIP score threshold of 1. Amongst the selected bins, only 54 were identified (28.57%) out of the 189, representing 14 unique metabolites (Figure 4.5-6). To establish the reliability of bins in representing their respective metabolite, CRS were calculated.

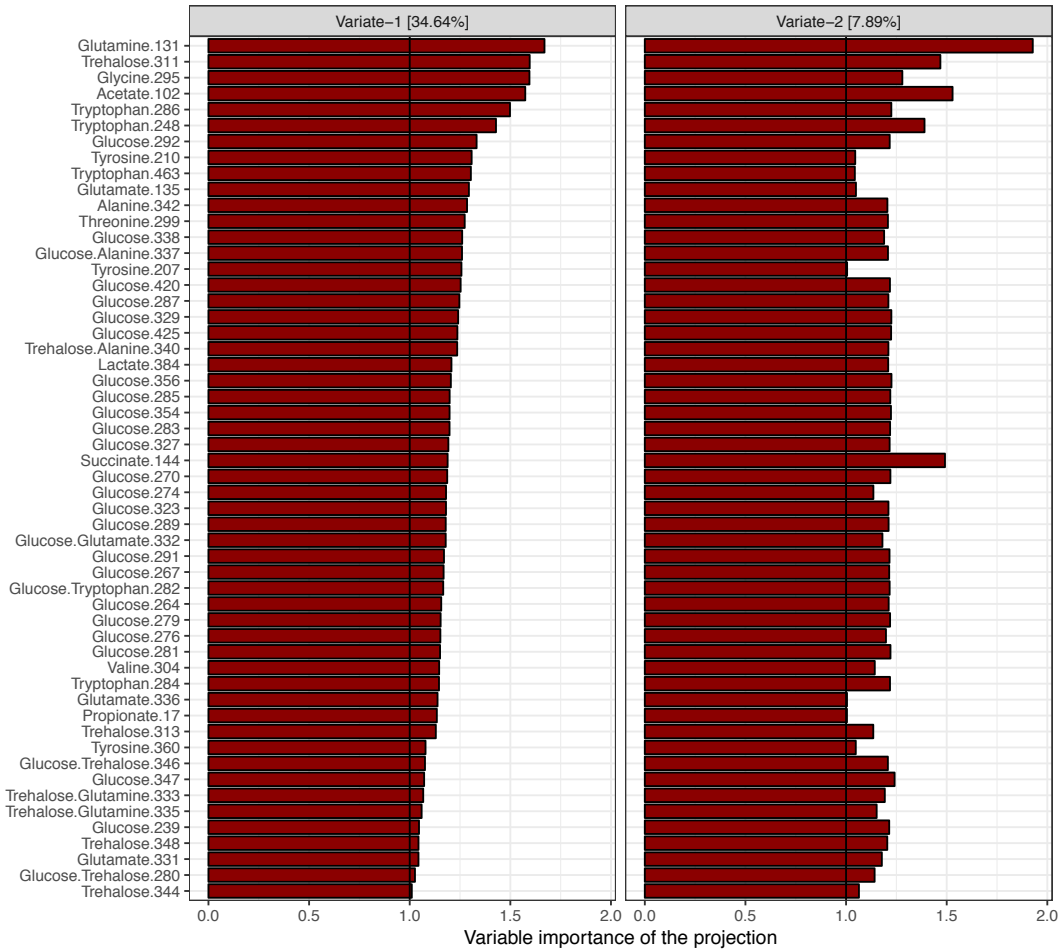


Figure 4.5-6: VIP scores from PLS-DA model of adult knock-down *An. gambiae* discriminating KD16 from KD17. Only identified bins scoring greater than one on both variate-1 and variate-2 were selected, totalling to 54 bins and representing 14 unique metabolites. Black line represents VIP score of 1.

CRS was used to calculate the passing score to be used as selection criterion. In order for a bin to be considered as a representative bin candidate, a score greater than 35.03% was required. CRS for the VIP selected bins were tabulated in Table 4.5-3, with representative bins selected as the highest scoring, non-overlapping (where applicable) bin.

Table 4.5-3: CRS for metabolites selected by VIP for KD16 and KD17. A minimum score of 35.03% was required for a bin to be considered as a representative of a particular metabolite. Non-overlapping peaks were selected where available.

Metabolite	Bin	CRS [%]	CRS > 35.03%	Rep	Metabolite	Bin	CRS [%]	CRS > 35.03%	Rep
Acetate	102	Singlet	NA	102	Glutamine	335*	30.05	×	-
Alanine	342	80.43	✓	342		333*	28.91	×	
	337*	79.80	✓			126*	10.00	×	
	340*	78.22	✓			131	7.90	×	
Glucose	285	96.89	✓	285	Glycine	295	Singlet	NA	295
	291	96.82	✓		Lactate	384	33.64	×	-
	279	96.81	✓		Propionate	15	73.98	✓	15
	270	96.76	✓			17	73.12	✓	
	283	96.75	✓			12	70.28	✓	
	267	96.73	✓			129	68.09	✓	
	264	96.70	✓			127	66.78	✓	
	289	96.69	✓			130	52.54	✓	
	281	96.64	✓		Succinate	144	Singlet	NA	144
	420	96.62	✓		Threonine	299	2.13	×	-
	329	96.58	✓		Trehalose	313	84.51	✓	313
	287	96.57	✓			280*	83.97	✓	
	354	96.56	✓			348	83.94	✓	
	425	96.45	✓			335*	82.37	✓	
	282*	96.42	✓			333*	81.50	✓	
	356	96.27	✓			346*	77.77	✓	
	292	96.22	✓			340*	70.96	✓	
	347	96.21	✓			344	70.93	✓	
	327	96.16	✓			311	63.05	✓	
	239	96.07	✓		Tryptophan	463	15.31	×	-
	323	95.56	✓			248	8.97	×	
	337*	95.09	✓			286	3.94	×	
	276	94.76	✓			284	2.58	×	
	338	94.66	✓			282*	0.99	×	
	332*	94.60	✓		Tyrosine	212	65.44	✓	212
	346*	94.47	✓			210	56.71	✓	
	280*	93.35	✓			207	56.14	✓	
	274	87.95	✓			360	-9.52	×	
Glutamate	140	22.39	×	-	Valine	304	-4.43	×	-
	138	22.16	×						
	135	19.58	×						
	331	5.93	×						
	332*	3.43	×						
	336	-5.40	×						
	126*	-6.23	×						

*: Overlapping bins

A shortlist of metabolites selected *via* CRS analysis are shown in Table 4.5-4. After the selection, the number of metabolites decreased from 14 to 8. Amongst the metabolites not selected were glutamate, glutamine, lactate, threonine, tryptophan and valine. The final list discriminating between KD16 and KD17 adults comprised amino acids (alanine, glycine and tyrosine); carboxylic acids (acetate, propionate and succinate), and saccharides (glucose and trehalose). Prior to MSEA, the influence of the selected metabolites in the differences between KD16 and KD17 were assessed.

Table 4.5-4: List of selected bins representing the metabolites influencing the metabolic differences between KD16 and KD17.

Class	Metabolite	Bin	Chemical shift [ppm]	KEGG code
Amino acids	Alanine	342	3.80	C00041
	Glycine	295	3.57	C00037
	Tyrosine	212	3.07	C00082
Carboxylic acids	Acetate	102	1.92	C00033
	Propionate	15	1.06	C00163
	Succinate	144	2.41	C00042
Saccharides	Glucose	285	3.49	C00031
	Trehalose	313	3.66	C01083

To observe the influence of the selected metabolites exclusively, raw data was filtered to include only the selected bins. PCA was then performed on the filtered dataset (Figure 4.5-7-A). PCA scores plot shows PC1 (50.16%) against PC3 (14.84%) accounting for a cumulative variance of 65.00%. A total of four components were required to explain the 95% variance in the data. Metabolic profiles of KD16 and KD17 explained by the selected metabolites show a difference along PC1 with minor augmentations on PC3, albeit with two samples distinct from the rest. The remaining samples could be easily separated. In order to assess the influences of selected metabolites in model building, a cross-validated PLS-DA model was built.

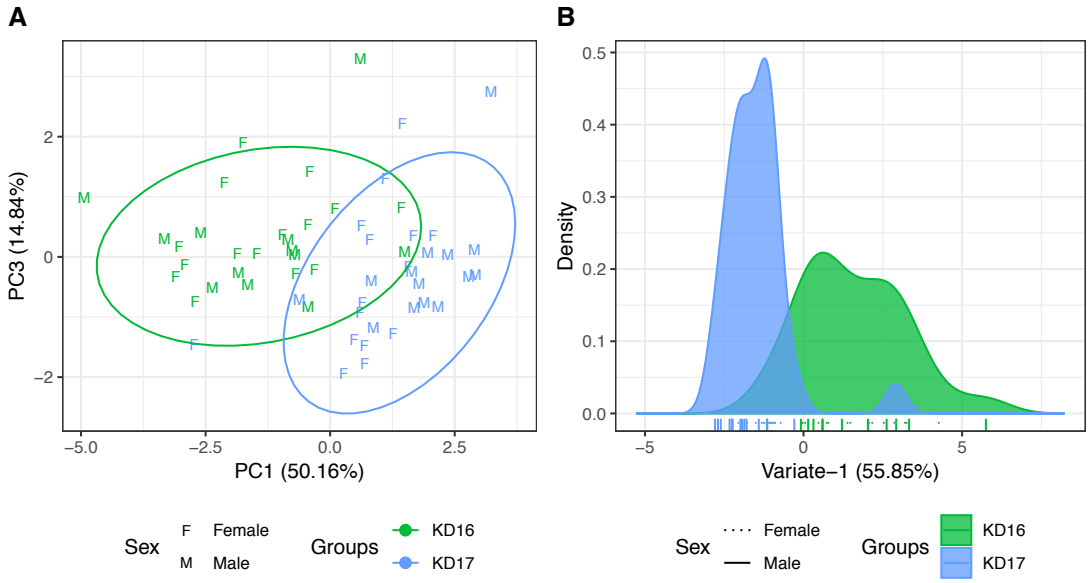


Figure 4.5-7: A) PCA scores plot (n_{KD16}=27, n_{KD17}=29) of knock-down adult data, showing PC1 (50.16%) and PC3 (14.84%) accounting for a total of 65.00% variance. A total of four components were required to explain 95% of the variance in the data. The ellipses represent 95% confidence region for each group. B) PLS-DA density plot of filtered adult knock-down *An. gambiae* data (n_{KD16}=27, n_{KD17}=29). Cross-validation was used to determine the single-variate (55.85% explained variance) complexity with 90.00% accuracy for the model. Each tick represents the samples from their respective group and each individual's sex.

A cross-validated PLS-DA model was built using the same filtered dataset optimised to a single-variate model with 90.00% accuracy (Appendix 11 for further metrics). The density plot (Figure 4.5-7-B) of the model shows a higher degree of variation in the KD16 samples compared to KD17 revealed by the broader features exhibited by KD16. A single KD17 sample is located within the KD16 group, showing the limitation and errors of the PLS-DA.

To further understand the differences between KD16 and KD17, metabolite levels were compared using a BH adjusted t-test (for detailed test statistics see Appendix 17) and were visualised by boxplots (Figure 4.5-8). In contrast to the pupae, all adult selected metabolites were significantly different between KD16 and KD17 post BH adjustment. The carboxylic acids propionate and acetate were found significantly higher in KD17, whereas succinate was significantly higher in KD16. Both saccharides (glucose and trehalose) were significantly higher in KD16. The amino acids (alanine and glycine) were also significantly higher in KD16, whereas tyrosine was significantly higher in KD17. These metabolites were subsequently analysed with MSEA, in order to probe the over-represented pathway by the selected metabolites.

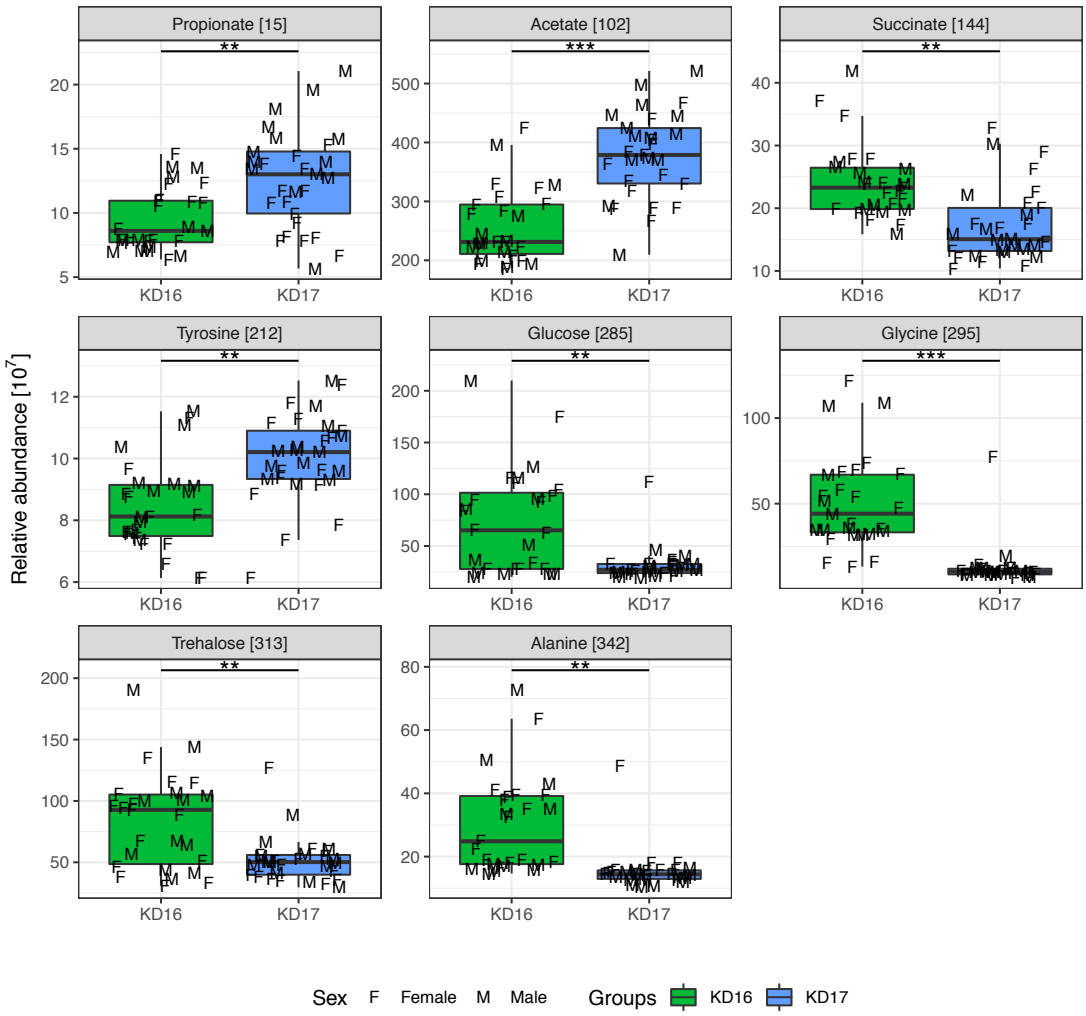


Figure 4.5-8: Boxplots of selected metabolites for knock-down *An. gambiae* adults (n_{KD16}=27, n_{KD17}=29). KD16, knock-down of Cyp4g16; KD17, knock-down of Cyp4g17. ** and *** denotes p-values less than 0.01 and 0.0001 respectively.

4.5.3 Metabolite set enrichment analysis

Selected metabolites from pupa and adult stages are summarised in Table 4.5-5, showing the metabolite levels of KD16 compared to KD17. A total of 11 metabolites were selected from the pupa PLS-DA model, whereas, only eight were selected from the adult PLS-DA model. From the selected metabolites in pupa, none were found to be significantly different. Within the selected metabolites, the CHC biosynthesis precursor valine was found to be lower in abundance compared to KD16 pupae. Although this difference was not statistically significant, it suggests KD17 exhibits a similar response to KD16, albeit to a lesser extent.

Table 4.5-5: Summary of selected metabolites of knock-down *An. gambiae* in pupa and adult. Levels are shown qualitatively for simplicity, although metabolites were compared quantitatively *via* t-test with BH p-value adjustment HSD. ↑, significantly higher; ↓, significantly lower, NS(↑); non-significantly higher, NS(↓); non-significantly lower, and square brackets represent BH-adjusted p-values.

	KD16 compared to KD17			
	Pupa selected		Adult selected	
Class	**Metabolite**	**Level**	**Level**	**Metabolite**
Amino acids	Alanine	NS (↓) [5.25 x10^{-1}]	↑ [1.37 x10^{-4}]	Alanine
	Glutamate	NS (↑) [3.81 x10^{-1}]		
	Glutamine	NS (↑) [1.12 x10^{-1}]		
			↑ [5.97 x10^{-7}]	Glycine
	Tryptophan	NS (↑) [5.25 x10^{-1}]		
			↓ [1.97 x10^{-4}]	Tyrosine
	Valine	NS (↑) [1.45 x10^{-1}]		
Carboxylic acids			↓ [1.81 x10^{-7}]	Acetate
	Formate	NS (↑) [4.53 x10^{-1}]		
	Propionate	NS (↑) [1.12 x10^{-1}]	↓ [5.59 x10^{-4}]	Propionate
	Pyruvate	NS (↓) [5.25 x10^{-1}]		
			↑ [1.70 x10^{-4}]	Succinate
Purines	Xanthine	NS (↓) [1.12 x10^{-1}]		
Saccharides	Glucose	NS (↑) [4.58 x10^{-1}]	↑ [2.91 x10^{-4}]	Glucose
	Trehalose	NS (↑) [1.45 x10^{-1}]	↑ [4.58 x10^{-4}]	Trehalose

Although some of these metabolites were not uniquely significant by univariate analysis, they were identified by PLS-DA as discriminating. Hence, these selected metabolites were used in MSEA with a pathway database curated from KEGG pathways (organism code: aga) using in-house scripts. Figure 4.5-9 shows the results of the MSEA for both pupae and adults. Both biosynthesis of amino acids and carbon metabolism are significantly over-represented pathways in knock-downs of both pupa and adults. Both of these pathways are over-represented in pupae with higher scores, caused by the higher number of amino acids in the pupal models (Table 4.5-6). The ABC transporters were also over-represented both in pupae and adults. Pathways over-represented uniquely include taurine & hypotaurine metabolism, nitrogen metabolism, glyoxylate and dicarboxylate metabolism, aminoacyl-tRNA biosynthesis, alanine, aspartate and glutamate metabolism, and 2-oxocarboxlic acid metabolism.

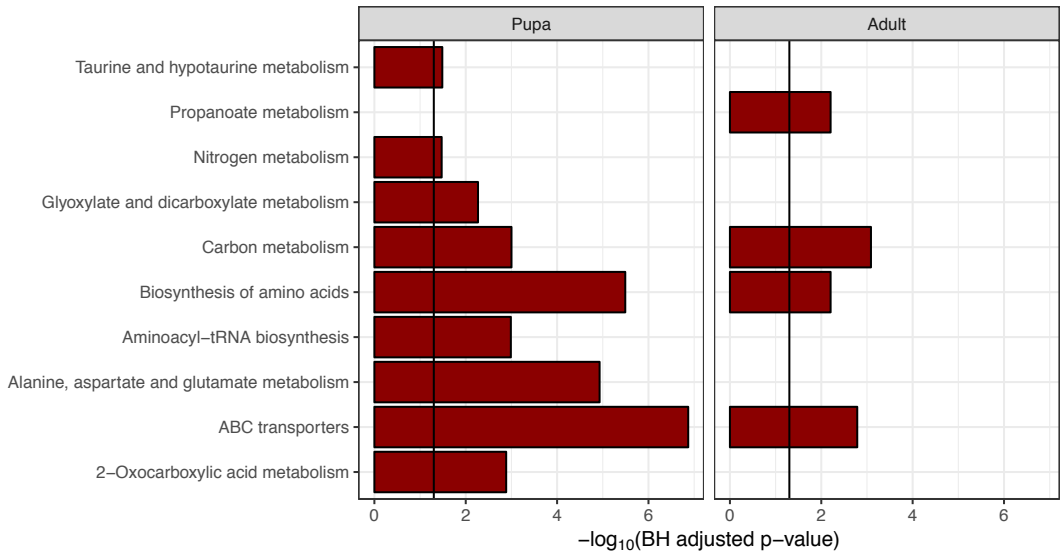


Figure 4.5-9: MSEA of KD16 and KD17 for pupae and adults. Black line represents p-value of 0.05. Pathways shown on the Figure are the most likely pathways to be altered between KD16 and KD17.

Table 4.5-6: Match table for identified MSEA pathways for two-way comparison of KD *An. gambiae* reporting; stage, raw & BH adjusted p-value, number of hits in the pathway and matching metabolites.

Pathway	Stage	Raw p-value	BH adjusted p-value	Hit/total (%)	Matches
2-Oxocarboxylic acid metabolism	Pupa	1.7×10^{-4}	1.3×10^{-3}	4/134 (2.99%)	Pyruvate, Glutamate, Tryptophan, Valine
ABC transporters	Pupa	3.0×10^{-9}	1.4×10^{-7}	6/182 (3.30%)	Glutamate, Glucose, Alanine, Glutamine, Valine, Trehalose
	Adult	8.8×10^{-5}	1.6×10^{-3}	4/182 (2.20%)	Glucose, Glycine, Alanine, Trehalose
Alanine, aspartate and glutamate metabolism	Pupa	7.7×10^{-7}	1.2×10^{-5}	4/29 (13.79%)	Pyruvate, Glutamate, Alanine, Glutamine
Aminoacyl-tRNA biosynthesis	Pupa	1.1×10^{-4}	1.0×10^{-3}	5/52 (9.62%)	Glutamate, Alanine, Glutamine, Tryptophan, Valine
Biosynthesis of amino acids	Pupa	1.4×10^{-7}	3.2×10^{-8}	6/128 (4.69%)	Pyruvate, Glutamate, Alanine, Glutamine, Tryptophan, Valine
	Adult	6.2×10^{-4}	6.3×10^{-3}	3/128 (2.34%)	Glycine, Alanine, Tyrosine
Carbon metabolism	Pupa	8.7×10^{-5}	1.0×10^{-3}	4/112 (3.57%)	Pyruvate, Glutamate, Alanine, Formate
	Adult	2.2×10^{-5}	8.2×10^{-4}	4/112 (3.57%)	Acetate, Glycine, Alanine, Succinate
Glyoxylate and dicarboxylate metabolism	Pupa	8.2×10^{-4}	5.4×10^{-3}	4/64 (6.25%)	Pyruvate, Glutamate, Formate, Glutamine
Nitrogen metabolism	Pupa	6.6×10^{-3}	3.4×10^{-2}	3/43 (6.98%)	Glutamate, Formate, Glutamine
Propionate metabolism	Adult	6.8×10^{-4}	6.3×10^{-3}	3/48 (6.25%)	Acetate, Succinate, Propionate
Taurine and hypotaurine metabolism	Pupa	5.7×10^{-3}	3.3×10^{-2}	3/22 (13.64%)	Pyruvate, Glutamate, Alanine

4.6 Chapter results summary

This chapter aimed to probe the metabolic differences exhibited by knock-downs of Cyp4g16 and Cyp4g17 species of *An. gambiae*. Cyp4g16 was shown to catalyse a critical step in the CHC biosynthesis and Cyp4g17 is hypothesised to catalyse the same reaction in a different location in oenocytes where both Cyp4g16 and Cyp4g17 enzymes reside [75]. Post metabolite selection, all models showed an approximate accuracy decrease of 5-10% hence, selected metabolites were less robust (Table 4.6-1). This indicates that some of the metabolites influential in the clustering is an unknown. Amongst the models, the worst performance was exhibited by the selected metabolite model for the pupal dataset, discriminating between all three groups with 69.70% accuracy. Overall, adult models outperformed pupa models.

Table 4.6-1: PLS-DA model variates, accuracy, VIP selection and selected metabolites.

Model strains	Stage	Data used	Model variates	Accuracy	VIP selection	Metabolites selected
KD16, KD17 and CONT	Pupa	All data	3	*75.76%	33 (25.78%)	8
	Adult		2	*82.35%	11 (11.00%)	6
	Pupa	Selected	2	*69.70%	NA	NA
	Adult	metabolites	2	*76.47%	NA	NA
KD16 and KD17	Pupa	All data	1	75.00%	41 (24.12%)	11
	Adult		2	100.00%	54 (28.57%)	8
	Pupa	Selected	2	75.00%	NA	NA
	Adult	metabolites	1	90.00%	NA	NA

* denotes average accuracy for a three-way model.

Selected metabolite comparisons from the three-way (KD16, KD17 & CONT) and two-way (KD16 & KD17) models comprised of amino acids, carboxylic acids, purines and saccharides (Table 4.6-2). In amino acids, glycine was higher in KD16 adults selected by the two-way model. Tryptophan was identified by all pupae models and was consistently higher than CONT. Valine was identified by the three-way model and was consistently higher than CONT in all comparisons. In the two-way comparison model, KD16 pupa showed higher levels of valine, but it was not statistically significant. In carboxylic acids, both acetate and succinate were selected from the two-way adult model and were higher in KD17 and KD16 respectively. Formate and pyruvate were selected by the two-way pupae model, and were higher in KD16 and KD17 pupae respectively, although, not significantly. For purines, xanthine was only selected from the two-way pupal model and showed elevated levels in KD17 albeit statistically significant. Lastly in saccharides, both glucose and trehalose were selected by all models. In the three-way comparison, glucose was significantly lower in KD17 adults compared to CONT. Meanwhile, in the two-way comparison, KD16 glucose levels were significantly higher in adults. This was also observed in trehalose levels with the addition of significantly higher levels in KD16 pupae and adults compared to CONT in three-way comparisons, but not in KD17.

Table 4.6-2: Metabolite level comparisons for knock-down comparisons in knock-down *An. gambiae*. Three-way comparison was made between KD16, KD17 and CONT and two-way comparison was made between KD16 and KD17. Arrows represent significant changes, NS (arrow) denotes non-significant change with mean abundance level, and square brackets report the BH-adjusted p-value.

		KD16 v KD17 v CONT				**KD16 v KD17**	
		KD16		**KD17**			
		Compared to CONT				**KD16 compared to KD17**	
Class	**Metabolite**	**Pupa**	**Adult**	**Pupa**	**Adult**	**Pupa**	**Adult**
Amino acids	Alanine		NS (↑) [6.31×10^{-2}]		↓ [6.31×10^{-3}]	NS (↓) [5.25×10^{-1}]	↑ [1.37×10^{-4}]
	Glutamate	NS (↑) [2.85×10^{-1}]	-	NS (↑) [5.23×10^{-2}]	-	NS (↑) [3.81×10^{-1}]	-
	Glutamine	-	-	-	-	NS (↑) [1.12×10^{-1}]	-
	Glycine	-	-	-	-	-	↑ [5.97×10^{-7}]
	Tryptophan	↑ [5.63×10^{-4}]	-	↑ [2.85×10^{-3}]	-	NS (↑) [5.25×10^{-1}]	-
	Tyrosine	NS (↓) [5.91×10^{-1}]	NS (↑) [5.91×10^{-1}]	NS (↓) [5.14×10^{-1}]	↑ [2.44×10^{-7}]	-	↓ [1.97×10^{-4}]
	Valine	↑ [5.35×10^{-4}]	↑ [5.71×10^{-3}]	NS (↑) [1.29×10^{-1}]	↑ [3.66×10^{-5}]	NS (↑) [1.45×10^{-1}]	-
Carboxylic acids	Acetate	-	-	-	-	-	↓ [1.81×10^{-7}]
	Formate	-	-	-	-	NS (↑) [4.53×10^{-1}]	-
	Propionate	↑ [2.66×10^{-2}]	-	NS (↑) [9.91×10^{-1}]	-	NS (↑) [1.12×10^{-1}]	↓ [5.59×10^{-4}]
	Pyruvate	-	-	-	-	NS (↓) [5.25×10^{-1}]	-
	Succinate	-	-	-	-	-	↑ [1.70×10^{-4}]
Purines	Xanthine	-	-	-	-	NS (↓) [1.12×10^{-1}]	-
Saccharides	Glucose	NS (↑) [2.28×10^{-1}]	NS (↑) [1.13×10^{-1}]	NS (↑) [9.24×10^{-2}]	↓ [3.04×10^{-3}]	NS (↑) [4.58×10^{-1}]	↑ [2.91×10^{-4}]
	Trehalose	↑ [5.69×10^{-3}]	↑ [3.48×10^{-2}]	NS (↑) [5.90×10^{-1}]	↓ [6.75×10^{-4}]	NS (↑) [1.45×10^{-1}]	↑ [4.58×10^{-4}]

MSEA results were summarised in Table 4.6-3. Propionate metabolism was the only differing pathway unique to adults. MSEA inferred differences in ABC transporters, carbon metabolism and biosynthesis of amino acids for both pupa and adult metabolic profiles in three-way comparison (KD16, KD17 and CONT). In the three-way comparison, 2-oxocarboxylic acid metabolism was only differing in pupae. MSEA comparing KD16 and KD17 identified ABC transporters, biosynthesis of amino acids and carbon metabolism differing both in pupae and adults. In the same comparison, pupae unique pathways were 2-oxocarboxylic acid metabolism, alanine, aspartate and glutamate metabolism, aminoacyl-tRNA biosynthesis, glyoxylate and dicarboxylate metabolism, nitrogen metabolism, and taurine & hypotaurine metabolism. Propionate metabolism was the only differing pathway unique to adults.

Table 4.6-3: MSEA summary for knock-down comparisons.

Over-represented pathways	KD16, KD17 & CONT	KD16 & KD17
2-oxocarboxylic acid metabolism	P	P
ABC transporters	C	C
Alanine, aspartate and glutamate metabolism		P
Aminoacyl-tRNA biosynthesis	C	P
Biosynthesis of amino acids	C	C
Carbon metabolism		C
Glyoxylate and dicarboxylate metabolism		P
Nitrogen metabolism		P
Propionate metabolism		A
Taurine and hypotaurine metabolism		P

A: unique to adults, C: common for both pupae and adults, P: unique to pupae.

4.7 Chapter discussion

Chapter 4 investigated the metabolite profile changes in the knock-downs of Cyp4g16 (KD16) and Cyp4g17 (KD17) in relation to their Gal4 homozygous controls (CONT), followed by a direct comparison of the two knock-downs (KD16 to KD17). Utilising a GAL4/UAS system in combination with NMR metabolomics allowed investigation of CHC biosynthesis through its precursor metabolites (valine, isoleucine and acetyl-CoA). Even though there has been extensive research performed on different aspects of CHC biosynthesis (done on different species), the metabolomics aspect remains unexplored. The knock-downs were hypothesised to increase the amount of the precursor metabolites to the CHC biosynthesis including valine, leucine, isoleucine and acetate as a proxy of the precursor acetyl-CoA.

A temporal expression of Cyp4g16 and Cyp4g17 was previously shown through transcriptomics by Balabanidou *et al* [75]. High mortality observed in KD16 pupae and KD17 adults during breeding in this project (Figure 4.2-1) are in line with the observations made by Lynd *et al* [188]. This further led to the expectation that KD16 (Cyp4g16 expressed highly in pupa) would demonstrate more differences compared to KD17 (Cyp4g17 expressed highly in adults). Upon knocking-down of Cyp4g16 and Cyp4g17, it was expected to create a shortage of final product HCs which is expected to create a build up of precursor metabolites.

In the analysis, valine was observed to be significantly higher in KD16 (pupae and adults) and in KD17 (adults) with respect to CONT. Although enzymes of Cyp4g16 and Cyp4g17 are thought to catalyse a similar reaction, they appear active at different developmental stages [188]. Significantly higher levels of valine indicate the activity of Cyp4g16 to be higher in

pupal stage (KD16:CONT = 1.33) and carries on to the early adult stage, whereas, Cyp4g17 shows greatest (as presented by valine level; KD17:CONT = 1.20) activity during early adult stage with respect to pupa. This is further supported by the similarities observed in the unsupervised multivariate analysis of pupal and adult datasets. The PCA scores plot highlights greater differences in KD17 compared to KD16 and CONT in adults. However, KD16 displays greater separation in pupae than CONT and KD17. The metabolites measured in this study do not separate according to sex. This indicates the affected HC biosynthesis pathways are not limited to sex characteristics specifically (e.g. specially leading to pheromone production). This further reveals the capability of both Cyp4g16 and Cyp4g17 in non-pheromone-specific HC production, specifically in the biosynthesis of n-alkanes, methyl-branched alkanes and 2-methyl-branched alkanes that partly form the cuticular layer.

Intriguingly, both precursors isoleucine and acetyl-CoA (acetate was used as a proxy) were neither identified as discriminating *via* PLS-DA nor found to be significantly higher compared to CONT. This further suggests that Cyp4g16 and Cyp4g17 are likely to have substrate specificity or higher preference for valine as the precursor rather than isoleucine, which synthesises HCs such as even chain length 2-methylalkanes [78], [86], [90] and methyl-branched alkanes [78], [86], [90]. The *D. melanogaster* homologue, Cyp4g1 (only decarbonylase) catalyses all decarbonylation of HCs. Interestingly, in the adult comparison, valine levels were significantly higher than CONT both in KD16 and KD17. The activity difference of Cyp4g16 and Cyp4g17 were observed phenotypically during mosquito rearing (Figure 4.2-1), although this is the first time this was shown with metabolite levels evidence from mosquito extracts. This metabolite level evidence supports the temporal activity of Cyp4g16 and Cyp4g17 shown by Lynd *et al* [212].

A recent study [186] has shown Cyp4g17 favours the decarbonylation of long chain methyl-branched alkanes (such as Me-hentriacontane) which utilise valine, leucine and isoleucine as precursors. Intriguingly, from the two-way PLS-DA models of KD16 & KD17, these precursors were not selected nor were they significantly different. In the metabolomics analysis of KD16 & KD17, valine levels were not different in the adults. Kefi *et al* created a *D. melanogaster* mutant where Cyp4g1 is knocked down and recovery was observed *via* introduction of Cyp4g16, Cyp4g17 and Cyp4g16/Cyp4g17 [186]. They have shown that decarbonylases Cyp4g1, Cyp4g16 and Cyp4g17 share a wide variety of common HC products. Furthermore, dimethyl-C45, dimethyl-C46 and dimethyl-C47 were only found in strains with Cyp4g16 and/or Cyp4g17 further supporting the specificity of these enzymes on methyl branched HCs.

Additionally, Kefi *et al* showed introduction of either Cyp4g16 or Cyp4g16/Cyp4g17 resulted in higher amounts of total CHC compared to Cyp4g17 introduced strains [186]. Taking into account, the rise in valine levels in early adults compounded with Kefti *et al* findings on preferences over methyl-branched HCs [186], further suggests Cyp4g16 has higher specificity for branched HCs which can be used for communication compared to Cyp4g17 where the majority HCs decarbonylated are more critical for waterproofing.

Since Cyp4g1 in *D. melanogaster* decarbonylates all of the HCs that Cyp4g16 and Cyp4g17 work on, it is entirely possible to speculate that, if both enzymes work on the same substrates there might be a competition between Cyp4g16 and Cyp4g17. The competition between Cyp4g16 and Cyp4g17 is in line with the findings of Kafi *et al* where reintroducing only Cyp4g16 or Cyp4g17 exclusively increased the HC biosynthesis [186]. Furthermore, the total CHC measured from the introductions of the Cyp4g16 and Cyp4g17 demonstrates that higher quantities of Cyp4g17 enzyme is required to decarbonylate the same amount of CHC as Cyp4g16 [186].

MSEA performed on the three-way comparison showed a consistent over-representation of the biosynthesis of amino acids, aminoacyl-tRNA biosynthesis and ABC transporters in the knock-down of Cyp4g16 and Cyp4g17. Biosynthesis of amino acids is consistent with the increased demand for valine to produce more CHC. Similarly, aminoacyl-tRNA biosynthesis is a downstream pathway of valine, leucine and isoleucine biosynthesis. Moreover, ABC transporters are used activating transportation of lipids (amongst other compounds) [112].

Acknowledging that the metabolomics work was done on early pupae and one day old adults, more apparent differences are potentially likely to be observed in older adults (3-5 days old). Early pupae and adults were chosen due to the lethal effects of knocking-down Cyp4g16 and Cyp4g17. KD16 pupae demonstrated high mortality in later pupae stages, whereas, KD17 demonstrated high mortality during and after emergence. For an appropriately comparable data, early pupa and adult samples were chosen. In order to increase the power of statistical modelling, higher numbers of samples could have been used although obtaining the necessary number of samples for such an experiment would require mosquito breeding on a very large scale. If such an experiment were to be carried out, I would further hypothesise KD17 adults would have complications in mating compared to CONT due to the reduced branched CHCs. Since branched CHCs can be used in communication a reduction in this type of CHC would lead to impairing of the communication between individuals hence effecting

mating. Alternatively, dead pupa could be investigated in order to understand the effects of these enzymes. Unfortunately, though, deconvolution of the metabolic changes contributed by death would be near impossible. It should also be noted that collected pupae would have a faulty CHC layer which would eventually cause them to drown and die. It is entirely possible that the current metabolism might be affected by the stress this has caused. In conclusion, through observation of valine levels in KD16 and KD17 supporting evidence was shown for their temporal activity as well as supporting evidence on KD17's decarbonylase activity.

Chapter 5

5 Understanding Pyrethroid Resistance using NMR Metabolomics

5.1 Introduction, chapter aims & objectives

Insecticide resistance is most commonly split into three types; knock-down resistance, metabolic resistance, and cuticular resistance [67]–[73]. Each one of these resistance types has been studies by various research groups using a variety of methods such as genomics, transcriptomics, and proteomics, but metabolomics studies of these comparisons are relatively lesser compared to the rest. Within the metabolomics studies, the most common focus is on the cuticular hydrocarbon (CHC) layer and even then, the data analysis does not usually make use of the multivariate statistical methods typically applied in metabolomics.

As covered in section 1.6.7, studies typically focus on the profiling of the CHC *via* GC-MS which has consistently showed the importance of *n*-alkanes (the most abundant constituent) and methyl-branched alkanes (second most abundant constituent) [84], [165], [166]. Although these studies are of great value, investigations into the metabolic comparison of resistant and susceptible strains with a focus on polar metabolites are lacking.

Any of these resistance mechanisms would produce a different response after coming into contact with an insecticide. Similarly, hypothesising these response mechanisms to have characteristic fingerprints on the metabolism in the absence of the insecticides would be logical. Hence, it is hypothesised that a resistance strain's metabolome would be distinctly different than that of a susceptible strain. By comparing the metabolic profiles of resistant and susceptible strains of *An. gambiae* and *Ae. aegypti*, these differences can be teased out. By doing so, externally known mechanisms or new undetected aspects of insecticide resistance can be attributed and explored.

5.2 Experimental Design

In this experiment, wild type mosquitoes were compared to obtain metabolic profiles of resistant and susceptible strains for *An. gambiae* and *Ae. aegypti*. All mosquitoes have been bred continuously in the LSTM insectary since their collection or acquisition from other laboratories. The resistant *An. gambiae* strain VK7 was collected from Burkina Faso and the susceptible *An. gambiae s*train N'Gusso was collected from Cameroon. Both strains have

been continuously bred for more than five years. For the *Ae. aegypti* comparisons, the resistant strain was collected from the Cayman Islands and the susceptible strain was collected from New Orleans. Both of these strains have also been continuously bred, in this instance for more than eight years. Table 5.2-1 shows the number of samples collected from the mosquitoes bred, successful NMR acquisitions, and spectra passing QC.

Table 5.2-1: Sample numbers used in statistical analysis post collection and QC. SUS: susceptible strain and RES: resistant strain.

	Samples collected				Spectra acquired				Quality control passed			
	An. gambiae		*Ae. aegypti*		*An. gambiae*		*Ae. aegypti*		*An. gambiae*		*Ae. aegypti*	
	SUS	RES	SUS	RES	SUS	RES	SUS	RES	SUS	RES	SUS	RES
Pupa	30	30	30	30	30	30	30	30	27	20	27	28
Adult	30	30	30	30	30	30	30	30	27	29	20	21

5.3 Metabolic profiling of wild type *An. gambiae* species VK7 (resistant) and N'gusso (susceptible)

5.3.1 Metabolite assignment

The bin tables for both pupa and adult *An. gambiae* were identical and consisted of a total of 496 bins. Out of the 496 bins, 114 where assigned to specific metabolites accounting for 22.98% of all bins, leaving 382 bins unidentified. Within the 114 assigned bins, 15 (13.16%) are overlapping bins. All 114 bins were attributed to a total of 21 unique metabolites (Table 5.3-1). The confidence levels of the metabolite assignments were scored according to the metabolite identification level (MSI). From the 21 metabolites, 16 (76.19%) were identified with MSI level 1, and 5 (23.81%) were identified with MSI level 2. Representative ^{1}H-NMR spectra for pupae and adults are shown in Appendix 18 and Appendix 19 respectively.

Table 5.3-1: Metabolite assignment table, with MSI level, KEGG compound code and classifications. See Appendix 1 for full NMR assignment table.

Classification	Metabolite	Metabolite identification level (MSI)	Unique	Overlap	Total	KEGG code
Alcohols	Methanol	Level 1	1	0	1	C00132
Amino acids	Alanine	Level 1	2	3	5	C00041
	Glutamate	Level 1	7	3	10	C00025
	Glutamine	Level 1	2	3	5	C00064
	Glycine	Level 1	1	0	1	C00037
	Isoleucine	Level 1	3	0	3	C00407
	Threonine	Level 1	4	0	4	C00188
	Tryptophan	Level 1	12	3	15	C00078
	Tyrosine	Level 1	12	2	14	C00082
	Valine	Level 1	5	0	5	C00183
Carboxylic acids	Acetate	Level 1	1	0	1	C00033
	Formate	Level 2a	1	0	1	C00058
	Fumarate	Level 2a	1	0	1	C00122
	Lactate	Level 1	4	0	4	C00186
	Propionate	Level 2b	6	0	6	C00163
	Pyruvate	Level 1	1	0	1	C00022
	Succinate	Level 1	1	0	1	C00042
Purines	Oxypurinol	Level 2b	1	0	1	C07599
	Xanthine	Level 2b	1	0	1	C00385
Saccharides	Glucose	Level 1	24	8	32	C00031
	Trehalose	Level 1	10	7	17	C01083

5.3.2 Metabolic profiling of pupae

5.3.2.1 Statistical analysis

In order to explore the major variances, PCA was performed. Figure 5.3-1-A shows PC1 (22.16%) against PC2 (20.75%) accounting for a total of 42.91% explained variance. In the PCA transformation, a total of 27 components were required to explain the 95% variance in the data. The variation spread along PC1 is similar between resistant and susceptible strains. The PCA scores plot shows a separation of resistant and susceptible strains along PC2. Using a supervised PLS-DA model, the metabolic profile observed on PC2 can be enhanced further, enabling metabolite selection in relation to the resistance status.

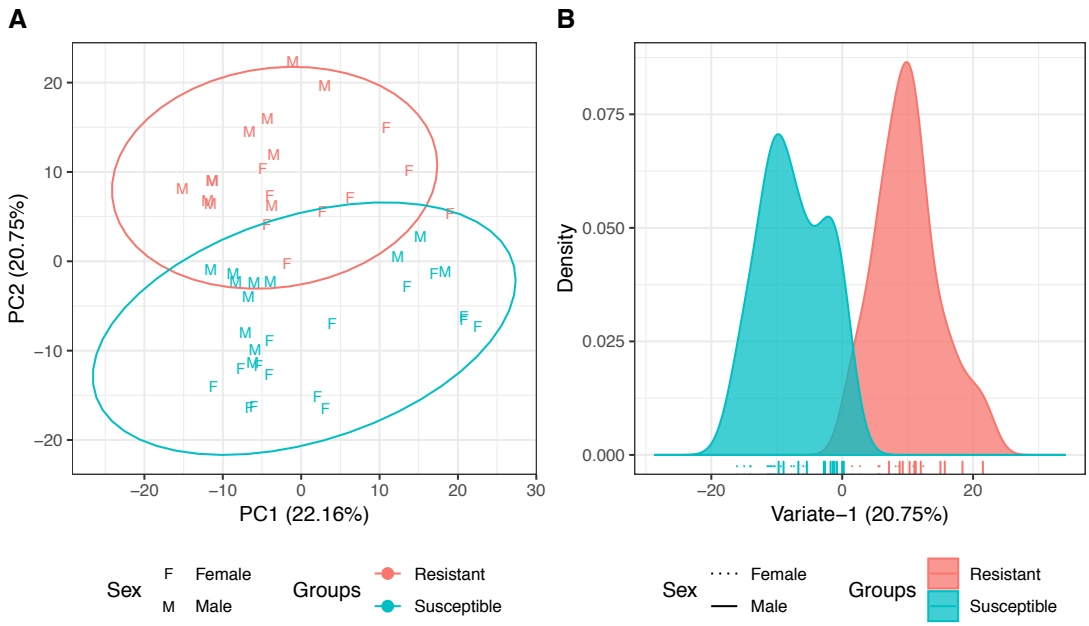


Figure 5.3-1: A) PCA scores plot of PC1(22.16%) and PC2 (20.75%) for wild type *An. gambiae* pupae demonstrating the overall similarity of the metabolic profiles of resistant and susceptible strains. Closer clusters indicate a higher degree of similarity. PC1 and PC2 accounts for a cumulative explained variance of 42.91%. A total of 27 components were required to explain 95% of the variance. Ellipses represent 95% confidence region of the groups individually. B) PLS-DA density plot of variate-1 discriminating between resistant and wild type *An. gambiae* pupae ($n_{Resistant}$=20, $n_{Susceptible}$=27). Model complexity of one-variate (20.75% explained variance) with accuracy of 64.29% was determined *via* cross-validation. Ticks represent samples from their respective groups.

A PLS-DA model (Figure 5.3-1-B) was built using cross-validation, optimal model complexity to be a single-variate model with 64.29% accuracy (Appendix 11 for further metrics). In order to probe the key metabolites of this discrimination, VIP scores were used.

5.3.2.2 Key metabolites

VIP scores were calculated for all 496 bins, including both identified and unidentified, using a passing threshold of 1 for bin filtering. Only 182 bins (identified and unidentified) scored greater than the VIP score of 1. Within the 182 bins, there were 52 identified bins representing a total of 18 metabolites (Figure 5.3-2). In order to select metabolites through a representative bin, correlation reliability score (CRS) was assessed.

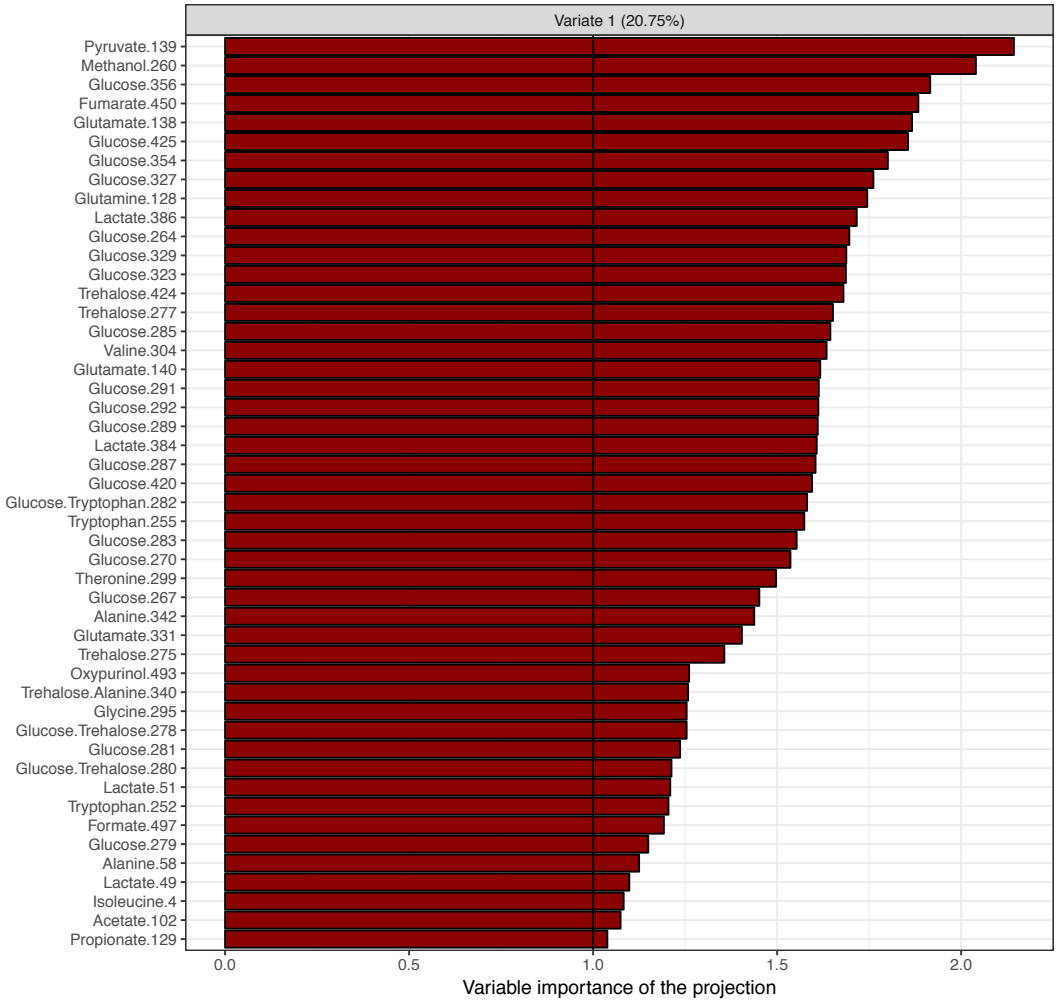


Figure 5.3-2: VIP scores of identified bins from the wild type pupae of *An. gambiae*. A total of 52 identified bins scored higher than the VIP threshold of one, representing 18 unique metabolites. Black line represents VIP score of one.

CRS was calculated for all 114 identified bins and filtered to show only the bins which scored higher than one on VIP scoring (Table 5.3-2). Using all CRS calculated, a passing score of 31.74% was set in order to accept a bin as a representative candidate. Only non-overlapping bins (where available) with the highest score amongst the candidates were selected.

Table 5.3-2: CRS for VIP selected bins from the wild type pupae *An. gambiae* PLS-DA model discriminating between resistant and susceptible strains. A passing score of 31.74% was calculated from the CRS scores of all 114 identified bins. Rep: representative bin.

Metabolite	Bin	CRS [%]	CRS > 31.74%	Rep	Metabolite	Bin	CRS [%]	CRS > 31.74%	Rep
Acetate	102	Singlet	NA	102	Glutamate	331	45.48	✓	331
Alanine	342	80.89	✓	342		138	43.40	✓	
	340*	78.94	✓			140	40.04	✓	
	58	72.74	✓		Glutamine	128	27.39	×	-
Formate	497	Singlet	NA	497	Glycine	295	Singlet	NA	295
Fumarate	450	Singlet	NA	450	Isoleucine	4	82.87	✓	4
Glucose	270	66.52	✓	270	Lactate	384	90.00	✓	384
	329	65.92	✓			51	87.32	✓	
	281	65.86	✓			49	85.74	✓	
	267	65.76	✓			386	77.89	✓	
	282*	65.72	✓		Methanol	260	Singlet	NA	260
	279	65.60	✓		Oxypurinol	493	Singlet	NA	493
	283	65.35	✓		Propionate	129	64.35	✓	129
	323	63.40	✓		Pyruvate	139	Singlet	NA	139
	264	62.99	✓		Threonine	299	13.82	×	299
	425	62.43	✓		Trehalose	275	75.96	✓	275
	327	62.26	✓			278*	70.47	✓	
	285	59.13	✓			280*	70.36	✓	
	287	58.87	✓			277	63.26	✓	
	354	56.79	✓			424	61.46	✓	
	420	56.77	✓			340*	20.26	×	
	292	56.62	✓		Tryptophan	252	36.00	✓	252
	356	56.51	✓			255	24.13	×	
	291	54.31	✓			282*	17.52	×	
	289	52.71	✓		Valine	304	34.01	×	-
	280*	5.16	×						
	278*	1.66	×						

*: Overlapping bins

Through VIP scoring, a total of 18 metabolites were selected out of 21 identified metabolites. VIP selected list was further shortened to 16 metabolites after applying CRS criteria (Table 5.3-3). Amongst the metabolites excluded from the selection were glutamine and valine. The 16 selected metabolites were comprised of five metabolite classes: alcohols, amino acids, carboxylic acids, purines and saccharides.

Table 5.3-3: Selected metabolites of PLS-DA model discriminating between wild type pupae of *An. gambiae* resistant and susceptible strains.

Class	Metabolite	Representative bin	Chemical shift [ppm]	KEGG code
Alcohol	Methanol	260	3.36	C00132
Amino acids	Alanine	342	3.80	C00041
	Glutamate	331	3.76	C00025
	Glycine	295	3.57	C00037
	Isoleucine	4	0.94	C00407
	Threonine	299	3.59	C00188
	Tryptophan	252	3.32	C00078
Carboxylic acids	Acetate	102	1.92	C00033
	Formate	497	8.46	C00058
	Fumarate	450	6.52	C00122
	Lactate	384	4.11	C00186
	Propionate	129	2.18	C00163
	Pyruvate	139	2.37	C00022
Purines	Oxypurinol	493	8.27	C07599
Saccharides	Glucose	270	3.42	C00031
	Trehalose	275	3.44	C01083

In order to determine the influence of these selected metabolites on discrimination of the groups, PCA was performed (Figure 5.3-3). PC1 (49.00%) and PC2 (14.82%) show a cumulative explained variance of 63.82%. A total of nine components were required to explain the 95% variance in the data. From the plot, it can be seen that the majority of the variation attributed to the differences between resistant and susceptible strains can be explained by PC1. PC2 shows a similar variation spread for both resistant and susceptible strains, with minor contribution to separation of resistance status compared to PC1.

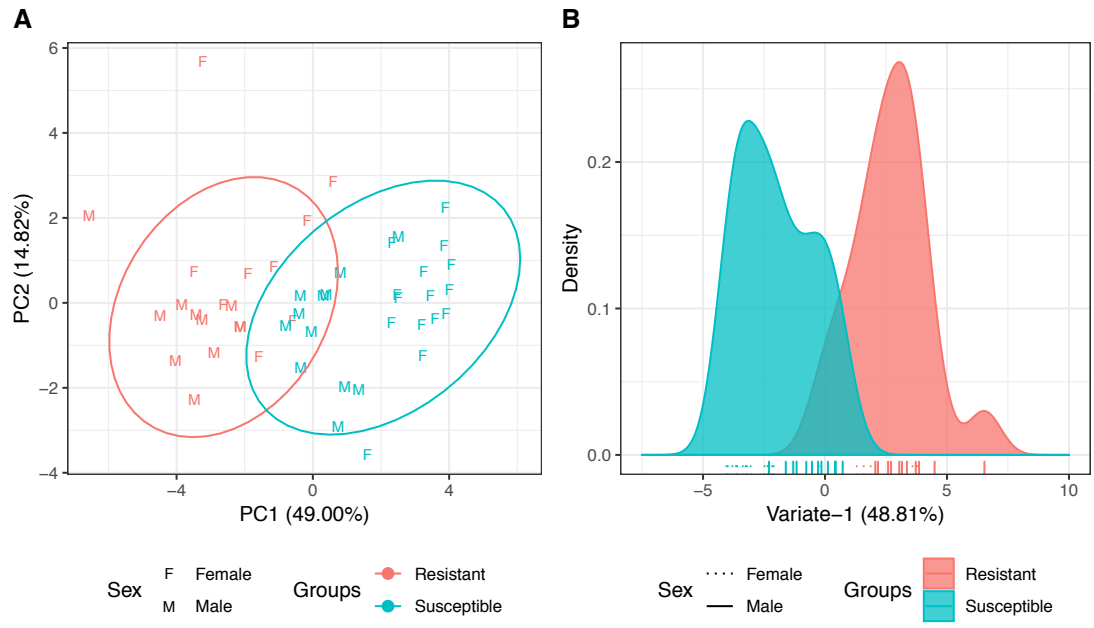


Figure 5.3-3: A) PCA scores plot of selected representative bins of *An. gambiae* pupae between resistant and susceptible groups ($n_{Resistant}$=20, $n_{Susceptible}$=27). A total of nine components were required to explain the 95% variance in the data. Ellipses represent the 95% confidence region. B) PLS-DA density plot of variate-1 discriminating between resistant and wild type *An. gambiae* pupae ($n_{Resistant}$=20, $n_{Susceptible}$=27) for selected representative bins. Single-variate (48.81% explained variance) model complexity was determined *via* cross-validation with 85.71% accuracy. Each tick under the density plot represents a sample.

A cross-validated PLS-DA model was built to discriminate between resistant and susceptible strains in order to determine the discriminative properties of the selected metabolites exclusively. Single-variate model complexity was determined through cross-validation and accuracy was calculated at 85.71% (Appendix 11 for further metrics). The resulting PLS-DA density plot shows the discrimination between two groups with only two resistant samples overlapping with susceptible samples.

Following the assessment of the selected metabolites' discriminatory properties, metabolite levels were compared in order to probe the metabolic phenotype of the resistant and susceptible groups. Metabolite levels were compared using a BH adjusted t-test (see Appendix 22 for detailed test statistics) and were presented in boxplots (Figure 5.3-4). All selected metabolites were significantly different between resistant and susceptible groups. More specifically, all carboxylic acids except lactate were significantly higher in resistant pupae. Significant carboxylic acids comprised of acetate, formate, fumarate, propionate and pyruvate. The only alcohol identified was methanol and it was found to be significantly higher in the resistant group. Within the amino acids, alanine, glycine, isoleucine and threonine were significantly lower in the resistant group, whereas glutamate and tryptophan were higher. The purine oxypurinol was found to be significantly higher in resistant species. Lastly,

in the saccharides, glucose was significantly lower in the resistant group while trehalose was higher.

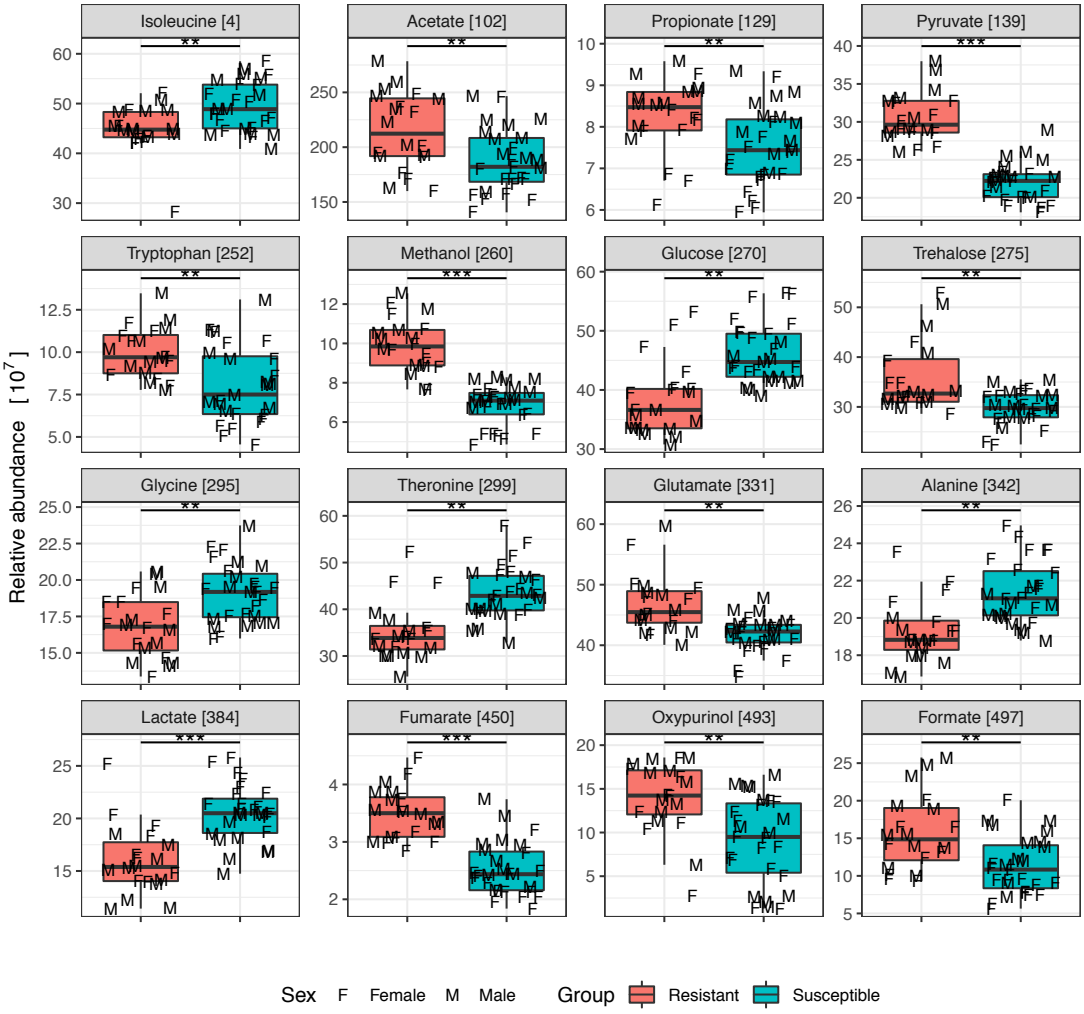


Figure 5.3-4: Boxplots for selected metabolites between wild type resistant and susceptible *An. gambiae* pupa. *, ** and *** denotes p-values less than 0.05, 0.01 and 0.0001 respectively.

5.3.3 Metabolic profiling of adults

5.3.3.1 Statistical analysis

The same approach that was taken for the pupae sample analysis was applied to the adult sample analysis in order to obtain a list of key metabolites representing the differences between resistant and susceptible strains. PCA (Figure 5.3-5-A) was performed on the adult data in order to observe the major variances. Differences between resistant and susceptible strains were explained clearest on PC2 (14.69%) and further enhanced by PC4 (4.25%). PC2 and PC4 accounted for a cumulative explained variance of 18.94%. Meanwhile, a total of 30 components were required in order to explain the 95% variance in the data. The majority of the metabolic profile difference can be observed along PC2, with a higher variation within

the resistant group compared to the susceptible group. PC4 enhances the metabolic profile difference observed on PC2, giving a clearer separation between two metabolic profiles. Using a cross-validated PLS-DA model, metabolites underlying the profile differences can be extracted.

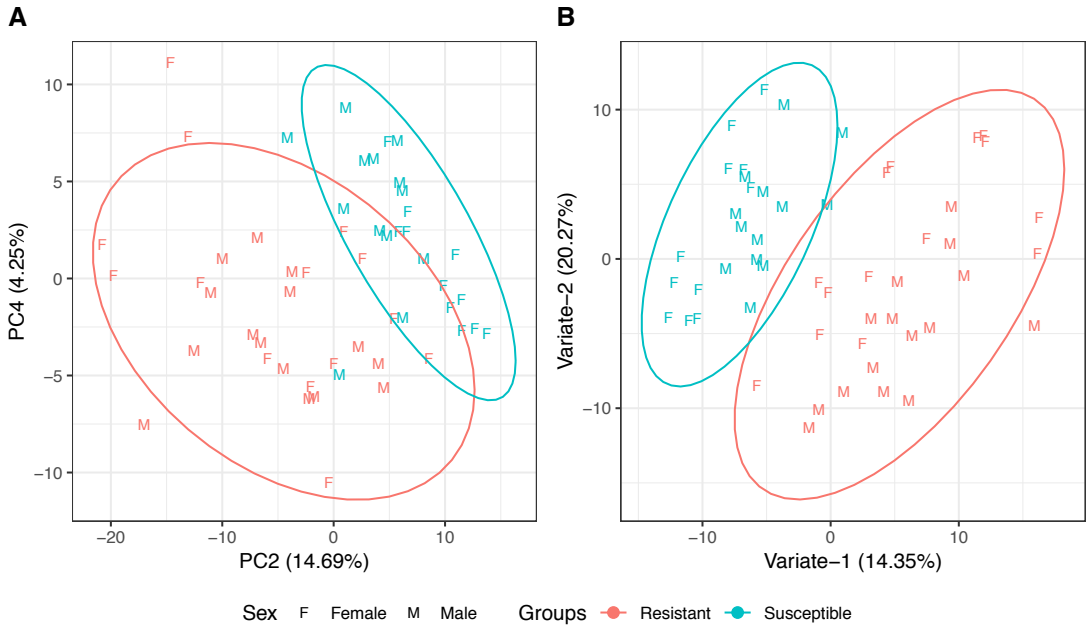


Figure 5.3-5: A) PCA scores plot of adult wild type *An. gambiae* ($n_{Resistant}$=29, $n_{Susceptible}$=27) showing PC2 (14.69%) and PC4 (4.25%) accounting for a total of 18.94% explained variance. A total of 30 components were required to account for 95% of the explained variance. Ellipses represent the 95% confidence region. B) PLS-DA scores plot discriminating between resistant and wild type *An. gambiae* adult ($n_{Resistant}$=29, $n_{Susceptible}$=24) along variate-1 and variate-2. Model complexity of two-variates was selected *via* cross-validation over 250 repetitions with 87.50% accuracy. Explained variance of each variate is reported in the brackets. Ellipses represent 95% confidence region.

A cross-validated PLS-DA model was built, accentuating the differences between resistant and susceptible strains. The optimal PLS-DA model complexity was determined to be a two-variate model with 87.50% accuracy (Appendix 11 for further metrics). The PLS-DA scores plot (Figure 5.3-5-B) shows a clear difference between profiles of resistant and susceptible species along the diagonal of variate-1 (14.35%) and variate-2 (20.27%). From the plot it is apparent that variate-1 accounts for the majority of the differences between the two metabolic profiles and variate-2 augments the observed difference.

5.3.3.2 Key metabolites

VIP scores were calculated for all 496 bins used in the model, including both identified and unidentified bins. A passing score threshold of 1 was applied on both variate-1 and variate-2 scores for VIP selection. Out of the 496 bins, only 157 bins (31.65%) scored higher than the threshold, which included both identified and unidentified bins. From the 157 bins scoring

higher than the threshold, only 30 bins were identified, representing 13 metabolites (Figure 5.3-6). The resulting list of bins was then used to select metabolites of influence in discriminating between resistant and susceptible strains. In order to generate this list, the representative qualities of the bins were assessed using CRS.

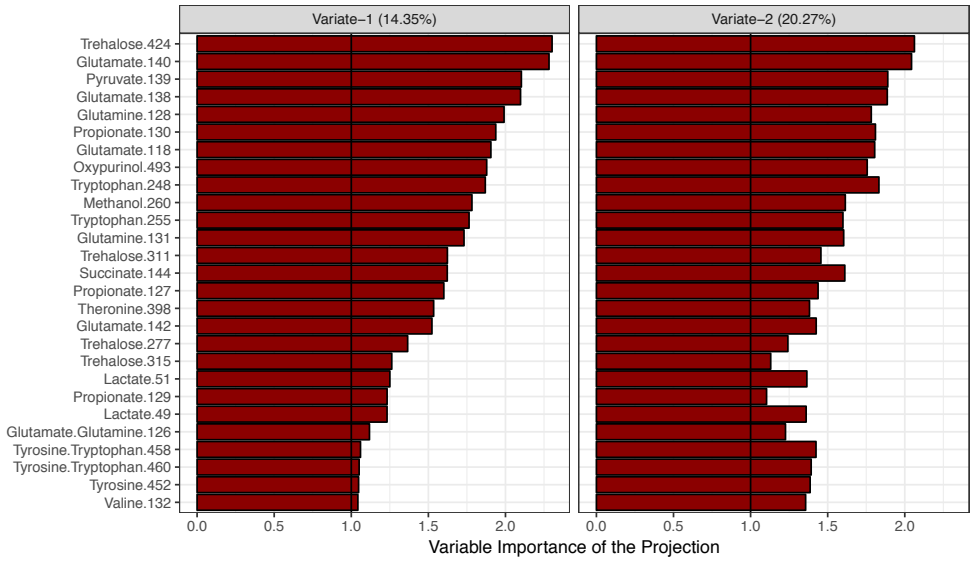


Figure 5.3-6: VIP scores from the PLS-DA model discriminating between resistant and susceptible adult *An. gambiae*. A total of 30 identified bins were selected representing 13 metabolites. Black line represent VIP score of 1.

The CRS was calculated for all identified bins and then used to calculate a passing score. The threshold for the adult *An. gambiae* dataset was 39.23%. Table 5.3-4 shows the CRS of 30 bins selected by VIP scoring criteria corresponding to 13 metabolites. In order to assess the representative properties of VIP selected bins, only bins scoring higher than CRS threshold were considered. Amongst the bins with sufficient score, only the highest scoring, non-overlapping (where applicable) were selected as metabolite representatives.

Table 5.3-4: CRS of bins selected from *via* VIP scores for adult *An. gambiae*. Only the highest scoring, non-overlapping bins (where applicable) over the CRS threshold were considered as representative bins. Rep: representative bin. * denotes overlapping bin.

Metabolites	Bin	CRS [%]	CRS > 39.23%	Rep	Metabolites	Bin	CRS [%]	CRS > 39.23%	Rep
Glutamate	140	21.34	×	-	Pyruvate	139	Singlet	NA	139
	118	20.98	×		Succinate	144	Singlet	NA	144
	138	20.69	×		Threonine	398	53.43	✓	398
	142	16.26	×		Trehalose	315	90.40	✓	315
	126*	-13.47	×			277	86.75	✓	
Glutamine	126*	26.62	×	-		311	79.26	✓	
	128	19.04	×			424	49.24	✓	
	131	-8.92	×		Tryptophan	248	16.79	×	-
Lactate	49	44.79	✓	49		255	14.18	×	
	51	43.94	✓			460*	13.76	×	
Methanol	260	Singlet	NA	260		458*	13.39	×	
Oxypurinol	493	Singlet	NA	493	Tyrosine	458*	77.21	✓	452
Propionate	129	59.25	✓	129		452	76.93	✓	
	127	48.26	✓			460*	76.90	✓	
	130	20.47	×		Valine	132	21.03	×	-

A metabolite shortlist (Table 5.3-5) was generated by selecting non-overlapping bins (where available) scoring higher than the 39.23% threshold. From the 13 metabolites identified by VIP, only nine were selected post-CRS. The four metabolites removed from the metabolite shortlist were: glutamate, glutamine, tryptophan, and valine. In order to determine the discriminative properties of these metabolites exclusively, adult data was filtered to only contain the selected bins representing their corresponding metabolites. These were then used in PCA to observe the major variance represented by the selected metabolites.

Table 5.3-5: Metabolite shortlist, curated from the VIP and CRS selection. Representative bins of selected metabolites.

Class	Metabolite	Representative bin	Chemical shift [ppm]	KEGG code
Alcohols	Methanol	260	3.36	C00132
Amino acids	Threonine	398	4.26	C00188
	Tyrosine	452	6.90	C00082
Carboxylic acids	Lactate	49	1.33	C00186
	Propionate	129	2.18	C00163
	Pyruvate	139	2.37	C00022
	Succinate	144	2.41	C00042
Purines	Oxypurinol	493	8.27	C07599
Saccharides	Trehalose	315	3.66	C01083

PCA (Figure 5.3-7-A) was performed on the filtered data to observe the major variances in the dataset when it only includes the selected metabolites. The PCA scores plot shows PC1 (34.87%) against PC3 (12.12%), accounting for a cumulative explained variance of 46.99%. A total of seven components were required to explain 95% of the variance. Although full separation of the profiles was not observed, an underlying structure is exhibited, suggesting the capability of selected metabolites in showing the differences between resistant and susceptible mosquitoes. A PLS-DA model was also built in order to determine the discriminatory properties of the selected metabolites represented by the bins.

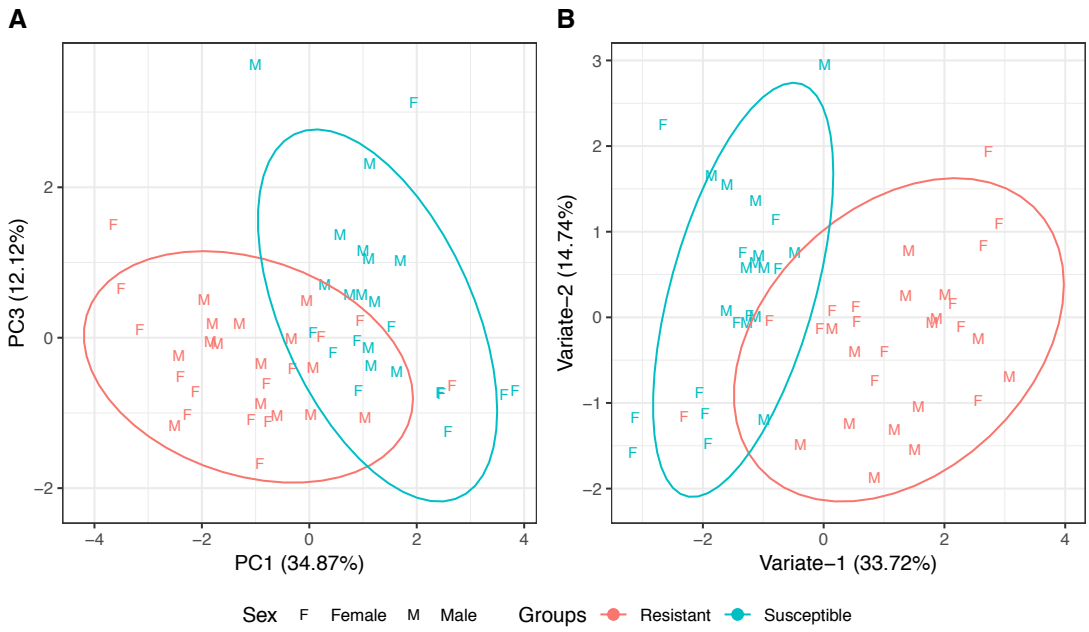


Figure 5.3-7: A) PCA of selected metabolites from the adult wild type *An. gambiae* PLS-DA model, discriminating between resistant and susceptible groups ($n_{Resistant}$=29, $n_{Susceptible}$=27). PC1 (34.87%) and PC3 (12.12%) account for a cumulative explained variance of 46.99%, while a total of seven components were required to explain the 95% variance in the data. Ellipses represent 95% confidence region. B) PLS-DA scores plot discriminating between resistant and wild type *An. gambiae* adult ($n_{Resistant}$=29, $n_{Susceptible}$=27) for selected representative bins. Model complexity of two variates was determined *via* cross-validation with 93.75% accuracy. Variate's explained variance is reported in the bracket. Ellipses represent 95% confidence region.

Discriminatory properties of the selected metabolites were assessed by a cross-validated PLS-DA model. Two-variate model complexity was determined *via* cross-validation with 93.75% accuracy (Appendix 11 for further metrics). Figure 5.3-7-B shows the PLS-DA scores plot for the filtered adult data. A separation of the resistant and susceptible groups can be seen along a diagonal of variate-1 and variate-2.

After establishing the discriminative properties of the selected metabolites, their levels were compared. Boxplots were used to represent the levels measured with BH adjusted t-test

applied (see Appendix 22 for detailed test statistics). Figure 5.3-8 shows that all selected metabolites were found to be significantly different (p-value < 0.05).

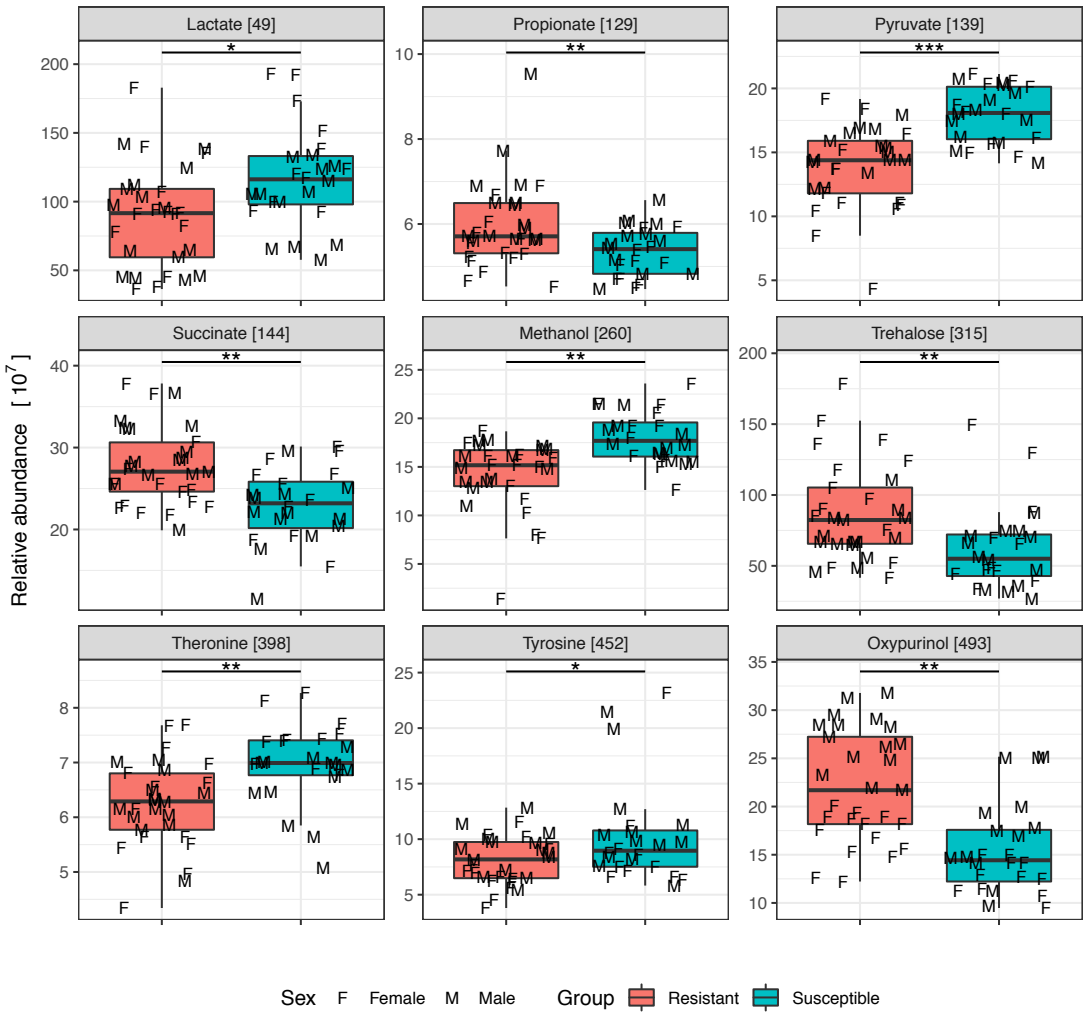


Figure 5.3-8: Boxplots for selected metabolites from adult wild type *An. gambiae*. * denotes p-value < 0.05, ** denotes p-value < 0.01 and *** denotes p-value < 0.0001.

5.3.4 Metabolite set enrichment analysis of pupal and adult stages

To summarise the key metabolite selection *via* statistical analyses, selected metabolites and their comparison to susceptible species were tabulated (Table 5.3-6). Selected metabolites were then used in MSEA to generate a list of pathways to represent the changes in resistant and susceptible species across pupae and adult stages.

Table 5.3-6: Selected metabolite level comparison summary of pupa and adult wild type *An. gambiae*. Arrows represent significant changes, and square brackets represent BH-adjusted p-values.

	Pupa		Adult	
Metabolite class	**Resistant : Susceptible**			
Alcohol	Methanol	↑ [8.17×10^{-9}]	↓ [1.90×10^{-4}]	Methanol
Amino acids	Alanine	↓ [2.23×10^{-4}]		
	Glutamate	↑ [7.22×10^{-4}]		
	Glycine	↓ [1.25×10^{-3}]		
	Isoleucine	↓ [2.89×10^{-3}]		
	Threonine	↓ [1.48×10^{-4}]	↓ [1.67×10^{-3}]	Threonine
	Tryptophan	↑ [1.67×10^{-3}]		
			↓ [4.28×10^{-2}]	Tyrosine
Carboxylic acids	Acetate	↑ [7.73×10^{-3}]		
	Formate	↑ [2.72×10^{-3}]		
	Fumarate	↑ [4.05×10^{-8}]		
	Lactate	↓ [6.12×10^{-5}]	↓ [1.07×10^{-2}]	Lactate
	Propionate	↑ [9.64×10^{-3}]	↑ [9.19×10^{-3}]	Propionate
	Pyruvate	↓ [3.79×10^{-10}]	↓ [9.50×10^{-6}]	Pyruvate
			↑ [1.03×10^{-3}]	Succinate
Purines	Oxypurinol	↑ [1.19×10^{-3}]	↑ [1.10×10^{-4}]	Oxypurinol
Sugar	Glucose	↓ [1.48×10^{-4}]		
	Trehalose	↑ [1.95×10^{-3}]	↑ [9.19×10^{-3}]	Trehalose

MSEA was perform on the selected metabolites of both pupae and adult wild type *An. gambiae* (Figure 5.3-9). MSEA was performed on selected metabolites and p-values were corrected with BH for multiple testing, resulting in a list of 15 unique pathways for pupae, two unique pathways for adult and five pathways in common for both pupae and adults (Table 5.3-7).

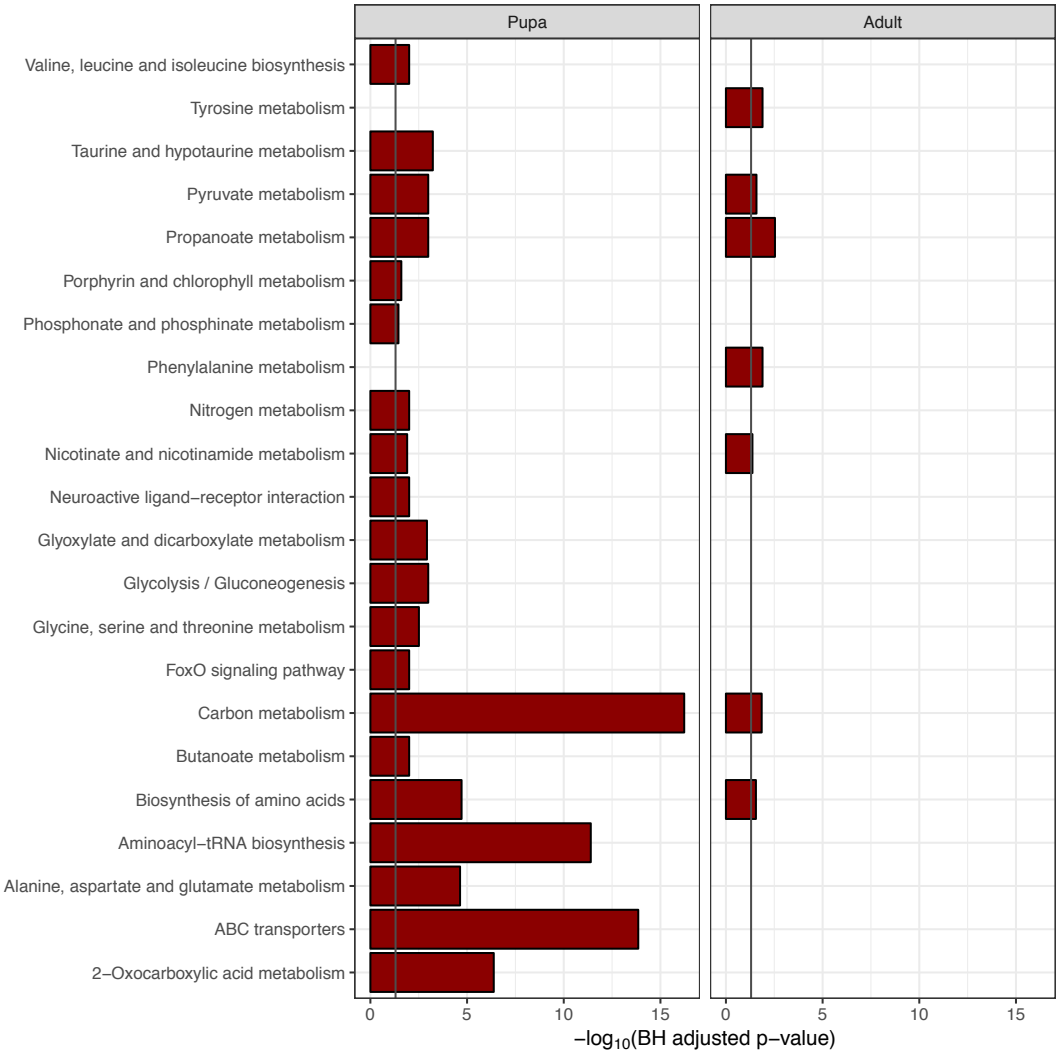


Figure 5.3-9: MSEA of selected metabolites of pupae and adults, showing over-represented pathways. Selected metabolites were derived from the PLS-DA models of pupae and adults discriminating between resistant and susceptible strains of *An. gambiae*. Pupae identified 20 significantly over-represented pathways and adults identified seven. Only five pathways were common in both pupae and adults. The black line represents p-value of 0.05.

Table 5.3-7: Wild type *An. gambiae* MSEA result details reporting Fisher's exact test raw & BH adjusted p-values, metabolite hit numbers per pathway and matching metabolites.

Pathway	Stage	Raw p-value	BH adjusted p-value	Hit/total (%)	Matches
2-Oxocarboxylic acid metabolism	Pupa	3.5×10^{-8}	4.2×10^{-7}	4/134 (2.99%)	Pyruvate, Glutamate, Tryptophan, Isoleucine
ABC transporters	Pupa	5.8×10^{-16}	1.4×10^{-14}	7/182 (3.85%)	Glutamate, Glucose, Glycine, Alanine, Threonine, Isoleucine, Trehalose
Alanine, aspartate and glutamate metabolism	Pupa	2.9×10^{-6}	2.3×10^{-5}	4/29 (13.79%)	Pyruvate, Glutamate, Alanine, Fumarate
Aminoacyl-tRNA biosynthesis	Pupa	2.5×10^{-13}	4.0×10^{-12}	6/52 (11.54%)	Glutamate, Glycine, Alanine, Tryptophan, Threonine, Isoleucine
Biosynthesis of amino acids	Pupa	2.0×10^{-6}	1.9×10^{-5}	7/128 (5.47%)	Pyruvate, Glutamate, Glycine, Alanine, Tryptophan, Threonine, Isoleucine
	Adult	5.0×10^{-3}	2.8×10^{-2}	3/128 (2.34%)	Pyruvate, Tyrosine, Threonine
Butanoate metabolism	Pupa	3.0×10^{-3}	9.9×10^{-3}	3/42 (7.14%)	Pyruvate, Glutamate, Fumarate
Carbon metabolism	Pupa	1.2×10^{-18}	5.9×10^{-17}	8/112 (7.14%)	Pyruvate, Glutamate, Acetate, Glycine, Alanine, Formate, Fumarate, Methanol
	Adult	1.7×10^{-3}	1.4×10^{-2}	3/112 (2.68%)	Pyruvate, Succinate, Methanol
FoxO signaling pathway	Pupa	3.5×10^{-3}	9.9×10^{-3}	2/5 (40%)	Glutamate, Glucose
Glycine, serine and threonine metabolism	Pupa	7.7×10^{-4}	3.1×10^{-3}	4/50 (8%)	Pyruvate, Glycine, Tryptophan, Threonine
Glycolysis / Gluconeogenesis	Pupa	1.9×10^{-4}	1.0×10^{-3}	4/31 (12.9%)	Pyruvate, Glucose, Acetate, Lactate
Glyoxylate and dicarboxylate metabolism	Pupa	2.7×10^{-4}	1.2×10^{-3}	4/64 (6.25%)	Pyruvate, Glutamate, Glycine, Formate
Neuroactive ligand-receptor interaction	Pupa	3.5×10^{-3}	9.9×10^{-3}	2/141 (1.42%)	Glutamate, Glycine
Nicotinate and nicotinamide metabolism	Pupa	4.7×10^{-3}	1.3×10^{-2}	3/55 (5.45%)	Pyruvate, Fumarate, Propionate
	Adult	8.8×10^{-3}	4.3×10^{-2}	3/55 (5.45%)	Pyruvate, Succinate, Propionate
Nitrogen metabolism	Pupa	3.5×10^{-3}	9.9×10^{-3}	2/43 (4.65%)	Glutamate, Formate
Phenylalanine metabolism	Adult	9.0×10^{-4}	1.3×10^{-2}	3/74 (4.05%)	Pyruvate, Succinate, Tyrosine
Phosphonate and phosphinate metabolism	Pupa	1.5×10^{-2}	3.5×10^{-2}	3/57 (5.26%)	Pyruvate, Acetate, Glycine
Porphyrin and chlorophyll metabolism	Pupa	9.9×10^{-3}	2.5×10^{-2}	3/141 (2.13%)	Glutamate, Glycine, Threonine
Propanoate metabolism	Pupa	2.1×10^{-4}	1.0×10^{-3}	4/48 (8.33%)	Acetate, Methanol, Propionate, Lactate
	Adult	8.6×10^{-5}	2.9×10^{-3}	4/48 (8.33%)	Succinate, Methanol, Propionate, Lactate
Pyruvate metabolism	Pupa	1.7×10^{-4}	1.0×10^{-3}	5/31 (16.13%)	Pyruvate, Acetate, Formate, Fumarate, Lactate
	Adult	3.9×10^{-3}	2.6×10^{-2}	3/31 (9.68%)	Pyruvate, Succinate, Lactate
Taurine and hypotaurine metabolism	Pupa	8.6×10^{-5}	5.9×10^{-4}	4/22 (18.18%)	Pyruvate, Glutamate, Acetate, Alanine
Tyrosine metabolism	Adult	1.1×10^{-3}	1.3×10^{-2}	3/79 (3.8%)	Pyruvate, Succinate, Tyrosine
Valine, leucine and isoleucine biosynthesis	Pupa	3.3×10^{-3}	9.9×10^{-3}	3/23 (13.04%)	Pyruvate, Threonine, Isoleucine

In the comparison of resistant and susceptible mosquitoes of *An. gambiae*, metabolic profile differences between resistant and susceptible strains were observed. From a metabolic profile point of view, these differences were much more pronounced in pupae compared to adults. PCA scores plots for pupae exhibited the resistance differences more clearly than those of adults. In addition, PCs demonstrating these differences also accounted for a higher cumulative explained variance in pupae (50.47%) compared to adults (18.94%). PLS-DA models also exhibited a similar performance: pupa model discriminated with a single-variate with 10.23% error, whereas the adult model required a 2-variate model to discriminate with 8.81% error rate. Upon performing MSEA, some common pathways were over-represented both in pupae and adult datasets. These were pyruvate metabolism, propanoate metabolism, nicotinate and nicotinamide metabolism, carbon metabolism and biosynthesis of amino acids. Furthermore, changes between resistant and susceptible strains were more pronounced in pupae than adults. Post-CRS metabolite selection yielded 16 metabolites for pupae while only nine metabolites were selected for adults. Although metabolic differences between resistant and susceptible strains were more pronounced in pupae, metabolite selection of adults has proved to be more informative. This was demonstrated by the performance of cross-validated PLS-DA models built with selected metabolites. In these models, pupae scored 85.71% accuracy. In comparison, the adult PLS-DA model scored 93.75% accuracy. This suggests that both models have similar performance in discriminating between resistant and susceptible species. Furthermore, it would appear that the adult model, which was built with fewer metabolites, could predict with higher accuracy. From the comparison of the metabolite levels it was found that levels of threonine (↓), propionate (↑), pyruvate (↓), oxypurinol (↑) and trehalose (↑) show significant changes and in the same direction for the pupae and adult dataset. On the contrary, methanol is the only metabolite which is significant in opposite directions between pupae (↑) and adults (↓).

5.4 Metabolic profiling of wild type *Ae. aegypti* species New Orleans (susceptible) and Cayman (resistant)

5.4.1 Metabolite assignment

Ae. aegypti metabolite assignment was consistent between pupae and adults comprising 513 bins in total (compared to 496 bins in *An. gambiae*). A total of 114 bins (same number as *An. gambiae*) were assigned out of 513, accounting for 22.22% of all bins, leaving 399 bins unidentified. The difference between *An. gambiae* and *Ae. aegypti* datasets were 17 unidentified bins. The metabolite assignment was identical for both resistant and susceptible

strains. From the 114 assigned bins, 15 (13.16%) were overlapping bins. A total of 21 unique metabolites were assigned from the 114 assigned bins (Table 5.4-1). From the 21 metabolites, 12 (57.14%) were identified with MSI level 1 and 9 were assigned (42.86%) with MSI level 2. Representative ^{1}H-NMR spectra for pupae and adults are shown in Appendix 20 and Appendix 21 respectively.

Table 5.4-1: Metabolite assignment table, with MSI level, KEGG compound code and classifications. See Appendix 1 for full assignment table.

Classification	Metabolite	Metabolite identification level (MSI)	Unique	Overlap	Total	KEGG code
Alcohols	Methanol	Level 1	1	0	1	C00132
Amino acids	Alanine	Level 1	2	3	5	C00041
	Glutamate	Level 1	7	3	10	C00025
	Glutamine	Level 1	2	3	5	C00064
	Glycine	Level 1	1	0	1	C00037
	Isoleucine	Level 1	3	0	3	C00407
	Threonine	Level 2	4	0	4	C00188
	Tryptophan	Level 1	12	3	15	C00078
	Tyrosine	Level 1	12	2	14	C00082
	Valine	Level 1	5	0	5	C00183
Carboxylic acids	Acetate	Level 1	1	0	1	C00033
	Formate	Level 2a	1	0	1	C00058
	Fumarate	Level 2a	1	0	1	C00122
	Lactate	Level 1	4	0	4	C00186
	Propionate	Level 2b	6	0	6	C00163
	Pyruvate	Level 1	1	0	1	C00022
	Succinate	Level 1	1	0	1	C00042
Purines	Oxypurinol	Level 2b	1	0	1	C07599
	Xanthine	Level 2b	1	0	1	C00385
Saccharides	Glucose	Level 1	24	8	32	C00031
	Trehalose	Level 1	10	7	17	C01083

5.4.2 Metabolic profiling of pupae

5.4.2.1 Statistical analysis

Major variances in the dataset were explored using PCA (Figure 5.4-1-A). PCA scores plot of PC1 (33.39%) against PC2 (14.82%) accounting for a cumulative variance of 48.21%. A total of 29 components were required in order to explain 95% variance in the data. The PCA plot shows a clear separation of resistant and susceptible species along a diagonal of PC1 and PC2. Observations on the plot also show subpopulations in each group which were attributed to sex. Using a cross-validated PLS-DA, differences between groups can be accentuated allowing for metabolite selection to be performed.

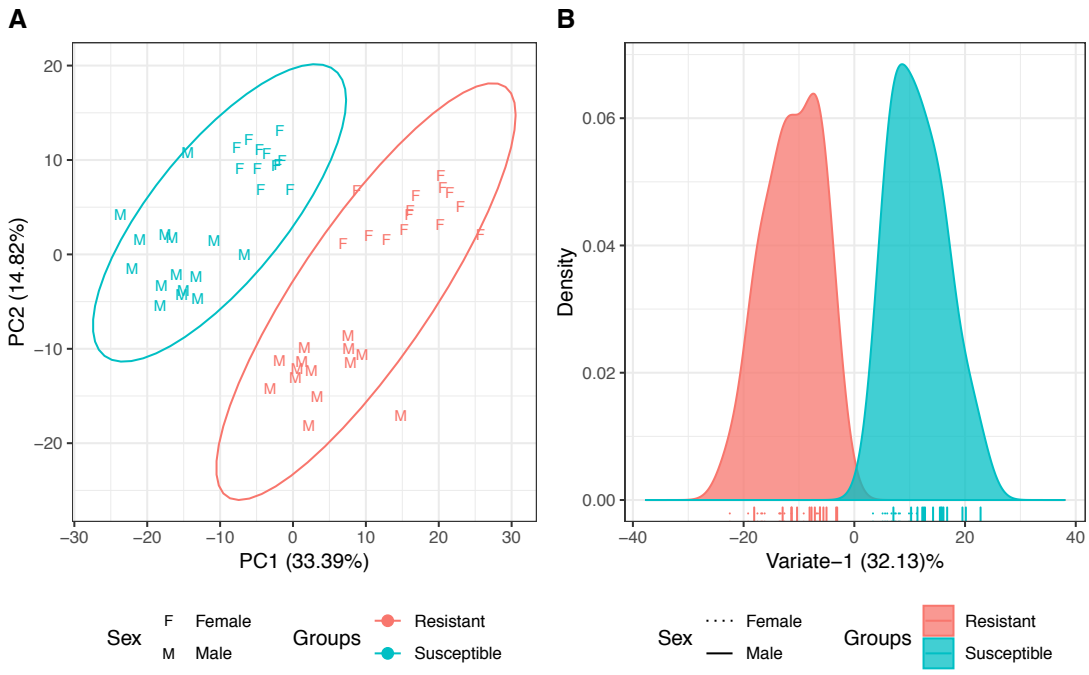


Figure 5.4-1: A) PCA scores plot PC1 (33.39%) against PC2 (14.82%) for wild type *Ae. aegypti* pupae ($n_{Resistnat}$=28, $n_{Susceptible}$=27) accounting for a cumulative variance of 48.21%. A high degree of metabolic profile difference is demonstrated by the two distinct clusters of resistant and susceptible groups. A total of 29 components were required to explain the 95% variance. The ellipses represent 95% confidence region. B) PLS-DA density plot for wild type *Ae. aegypti* pupae ($n_{Resistnat}$=28, $n_{Susceptible}$=27) PLSDA. Single-variate model (32.13% explained variance) complexity was determined using cross-validation and model accuracy was calculated at 100%. Each tick represents a sample from its respective group and its sex.

Using a cross-validated PLS-DA model (Figure 5.4-1-B), sex variance observed in the PCA were supressed and the differences between resistant and susceptible species were enhanced. Single-variate optimal model complexity was determined through cross-validation with 100% accuracy (Appendix 11 for further metrics). PLS-DA density plot of variate-1 shows clear discrimination of the two groups. In order to identify metabolites influencing the discrimination in the model, VIP scores were calculated for the bins.

5.4.2.2 Key metabolites

VIP scores were calculated to identify the bins most influential in discriminating between resistant and susceptible strains of *Ae. aegypti* pupae. A total of 513 VIP scores were calculated including both identified and unidentified bins. A VIP score threshold of 1 was applied for selection, where 245 bins scored above. Within the 245 bins, only 60 were identified and these were attributed to 15 unique metabolites. Therefore, in order to select the best representative bin for a metabolite, CRS was applied.

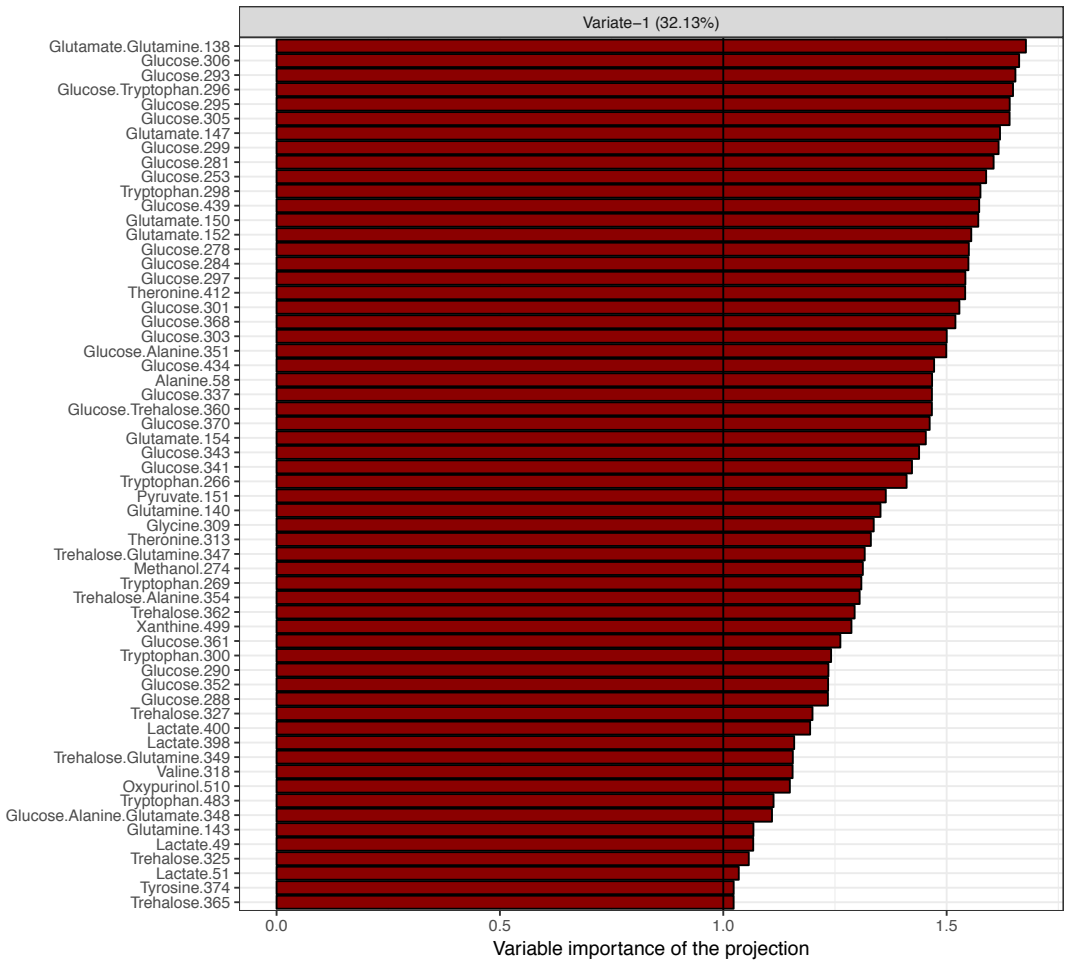


Figure 5.4-2: Identified bins scoring higher than 1 on VIP scoring from PLS-DA model discriminating between pupae of wild type *Ae. aegypti* resistant and susceptible strains. Black line represents VIP score of 1.

Using all CRS calculated from identified bins (114), a passing score of 42.82% was calculated. CRS scores of bins selected from VIP were checked against the passing score (Table 5.4-2). A selection amongst bins scoring higher than the passing score was made. Only the highest scoring, non-overlapping bins (where applicable) were selected.

Table 5.4-2: CRS of bins selected from *via* VIP scores for *Ae. aegypti* pupae. Only the highest scoring, non-overlapping bins (where applicable) over the CRS threshold of 42.82% were considered as representative bins.

Metabolite	Bin	CRS [%]	CRS > 42.82%	Rep	Metabolite	Bin	CRS [%]	CRS > 42.82%	Rep
Alanine	351*	79.77	✓	58	Glutamine	349*	9.41	×	-
	354*	77.51	✓			347*	9.07	×	
	348*	73.04	✓			143	8.08	×	
	58	66.19	✓			138*	0.92	×	
Glucose	284	87.41	✓	284		140	-0.28	×	
	368	87.30	✓		Glycine	309	Singlet	NA	309
	370	87.13	✓		Lactate	398	90.09	✓	398
	281	86.53	✓			49	86.54	✓	
	303	86.41	✓			51	85.31	✓	
	296*	86.39	✓			400	78.49	✓	
	278	86.35	✓		Methanol	274	Singlet	NA	274
	439	86.32	✓		Oxypurinol	510	Singlet	NA	510
	360*	86.18	✓		Pyruvate	151	Singlet	NA	151
	306	86.09	✓		Threonine	313	18.17	×	-
	299	85.98	✓			412	-4.46	×	
	253	85.94	✓		Trehalose	365	79.92	✓	365
	343	85.92	✓			325	79.42	✓	
	341	85.75	✓			327	73.52	✓	
	337	85.74	✓			362	70.75	✓	
	305	85.71	✓			360*	67.07	✓	
	434	85.66	✓			349*	66.61	✓	
	295	85.56	✓			347*	57.38	✓	
	301	85.53	✓			354*	45.27	×	
	297	85.46	✓		Tryptophan	483	29.31	×	-
	293	85.12	✓			298	21.71	×	
	351*	85.04	✓			300	20.56	×	
	352	83.53	✓			266	20.18	×	
	361	81.85	✓			296*	13.09	×	
	290	81.32	✓			269	7.95	×	
	288	80.75	✓		Tyrosine	374	62.67	✓	374
	348*	68.48	✓		Valine	318	18.56	×	-
Glutamate	150	36.59	×	-	Xanthine	499	Singlet	NA	499
	152	36.27	×						
	147	35.36	×						
	154	33.75	×						
	138*	22.55	×						
	348*	-0.98	×						

*: Overlapping bin

A metabolite shortlist was tabulated from the selected bins from CRS. Prior to CRS, 15 metabolites were selected *via* VIP. Using CRS, the metabolites glutamine, glutamate, threonine, tryptophan and valine were excluded, resulting in a list of 10 metabolites. Selected metabolites were comprised of five classes: alcohols, amino acids, carboxylic acids, purines and saccharides.

Table 5.4-3: Metabolite shortlist for wild type pupae of *Ae. aegypti* discrimination between susceptible and resistant strains.

Class	Metabolite	Representative bin	Chemical shift [ppm]	KEGG code
Alcohols	Methanol	274	3.36	C00132
Amino acids	Alanine	58	1.48	C00041
	Glycine	309	3.57	C00037
	Tyrosine	374	3.94	C00082
Carboxylic acids	Lactate	398	4.11	C00186
	Pyruvate	151	2.37	C00022
Purines	Oxypurinol	510	8.27	C07599
	Xanthine	499	7.89	C00385
Saccharides	Glucose	284	3.42	C00031
	Trehalose	365	3.87	C01083

Prior to MSEA, the major variances represented by the selected metabolites were evaluated using PCA. NMR data was filtered to include only the selected metabolites to perform PCA (Figure 5.4-3-A). PC1 (56.55%) against PC2 (15.17%) explains a cumulative variance of 71.72% and a total of 7 components were required to explain the 95% variation in the metabolic profile of resistant and susceptible species. Separation between resistant and susceptible species with no overlap can be observed along PC1. Both groups exhibit similar internal group variation demonstrated by the spread along PC1. Further discriminatory properties of the selected metabolites were assessed *via* cross-validated PLS-DA model.

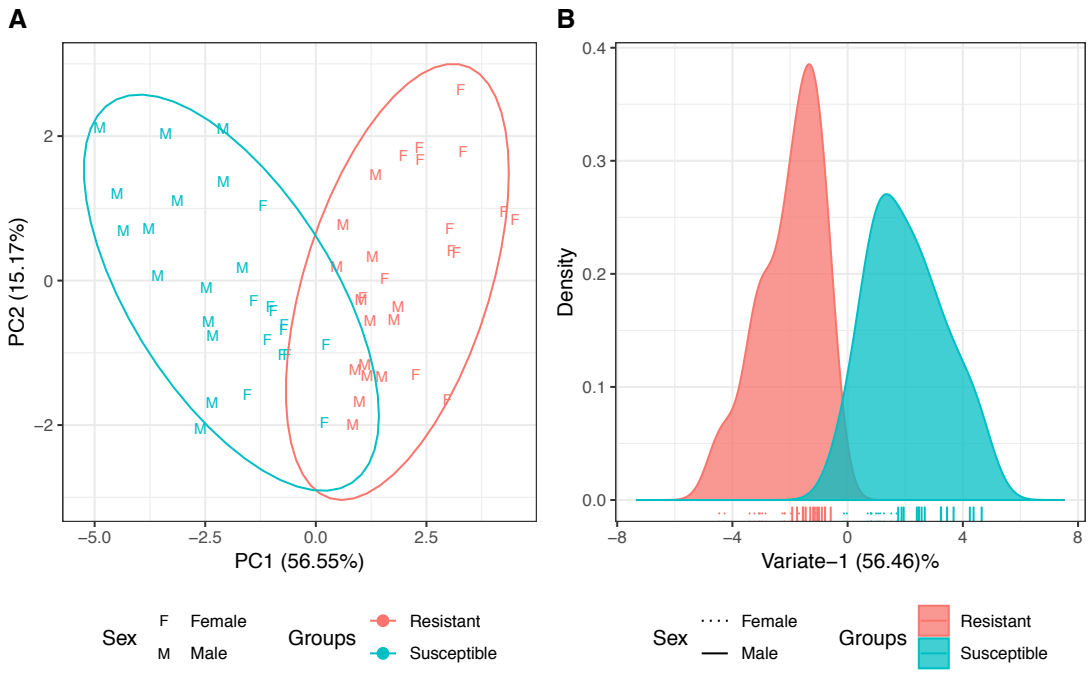


Figure 5.4-3: A) PCA scores plot of PC1 (56.55%) against PC2 (15.17%) for wild type *Ae*. *aegypti* pupae ($n_{Resistant}$=28, $n_{Susceptible}$=27). PC1 and PC2 explain a cumulative variance of 71.72% in the metabolic profile between resistant and susceptible species. A total of seven components were required to explain 95% variation in the data. Ellipses represent the 95% confidence region. B) PLS-DA density plot for wild type *Ae*. *aegypti* pupae ($n_{Resistant}$=28, $n_{Susceptible}$=27) for selected representative bins. Single variate (56.46% explained variance) model complexity was determined by cross-validation with 87.50% accuracy. Each tick represents a sample and its sex from their groups.

A cross-validated PLS-DA model was built to discriminate between resistant and susceptible strains. Optimal model complexity was determined to be a single-variate model with 87.50% accuracy (Appendix 11 for further metrics). The PLS-DA density plot (Figure 5.4-3-B) shows clear discrimination between resistant and susceptible strains. To further probe the metabolite level changes, selected metabolites were investigated.

Using a BH adjusted t-test, selected metabolites were compared between resistant and susceptible strains (see Appendix 22 for detailed test statistics). Metabolite levels were visualised *via* boxplots (Figure 5.4-4). Upon BH adjustment, all metabolites were significantly different between resistant and susceptible strains. In the case of alcohols, methanol was found to be lower in resistant strains. For the amino acids, alanine, glycine and tyrosine were significantly higher in resistant strains. For carboxylic acids, resistant species exhibited higher levels of lactate while showing lower levels of pyruvate. In purines, lower levels of oxypurinol were observed in resistant species while xanthine levels were higher. Finally, for saccharides, both glucose and trehalose were higher in resistant species.

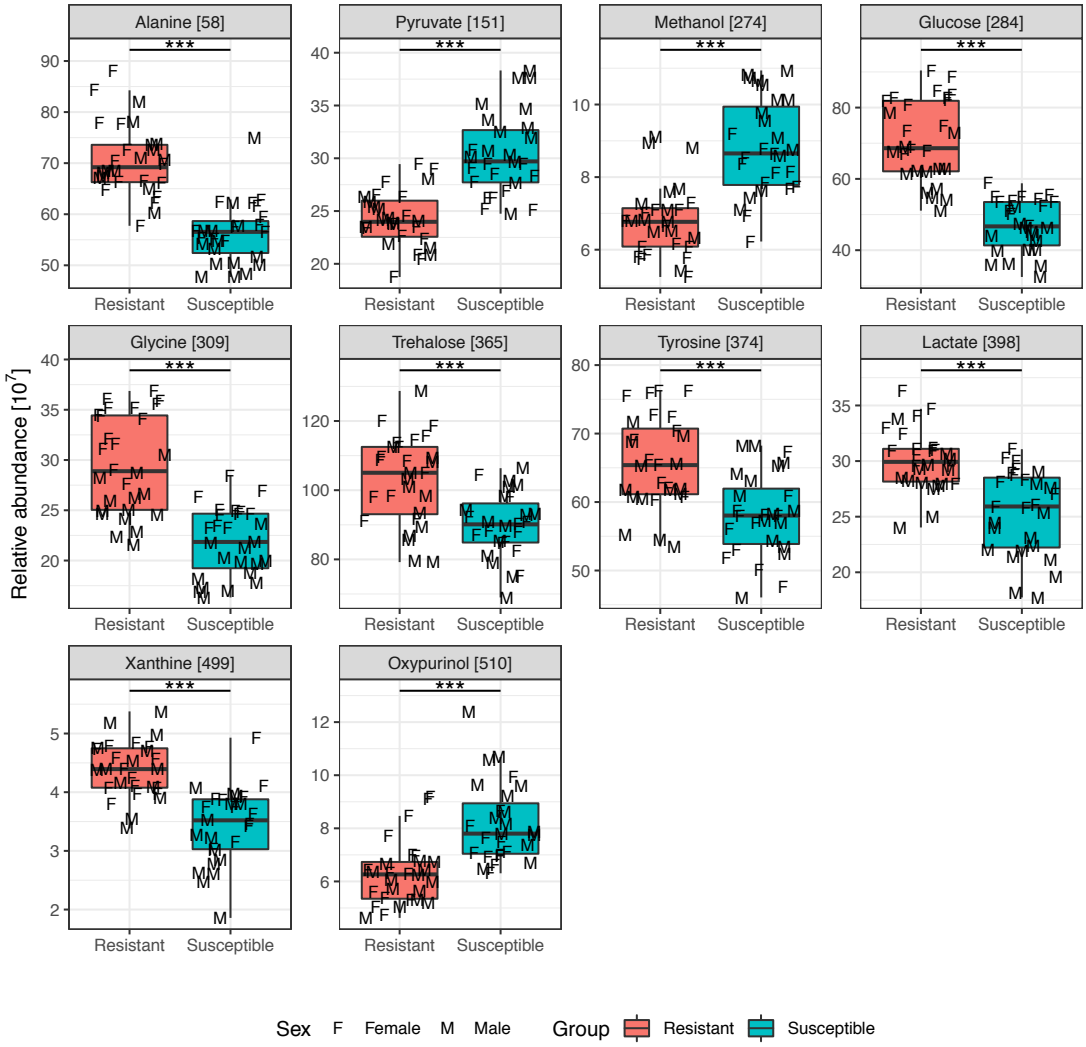


Figure 5.4-4: Boxplots of selected metabolites between wild type resistant and susceptible *Ae. aegypti.* *, ** and *** denotes p-value less than 0.05, 0.01 and 0.0001 respectively.

5.4.3 Metabolic profiling of adults

5.4.3.1 Statistical analysis

In order to compare the pupal and adult phases, the adult data was analysed with the same approach as the pupal analysis. Using PCA (Figure 5.4-5-A), major variances in the adult dataset were observed. Resistant and susceptible strains of adult *Ae. aegypti* were clearly separated along a diagonal on PC1 (29.14%) and PC2 (13.17%) accounting for a cumulative explained variance of 42.31%. A total of 25 components were required to explain the 95% variance in the data. In order to select metabolites of interest between resistant and susceptible strains, metabolite differences were accentuated *via* PLS-DA modelling.

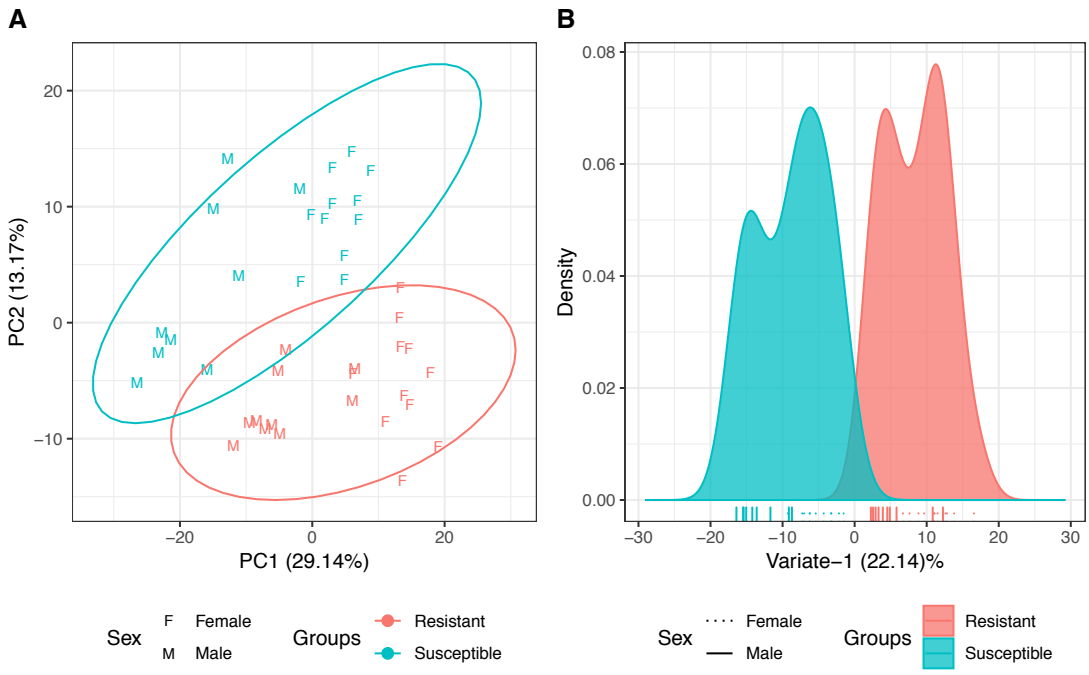


Figure 5.4-5: A) PCA scores plot for wild type *Ae. aegypti* adult ($n_{Resistant}$=21, $n_{Susceptible}$=20). PC1 (29.14%) and PC2 (13.17%) account for a cumulative variance of 42.31%, meanwhile, 25 components were required to explain 95% of the variance. Ellipses represent the 95% confidence region. B) PLS-DA density plot discriminating between resistant and susceptible strains of wild type *Ae. aegypti* adult ($n_{Resistant}$=21, $n_{Susceptible}$=20). Single variate model (22.14% explained variance) complexity was determined *via* cross-validation 100% accuracy. Two subpopulations in each group can be observed due to sex as shown by the different line types of the tick marks. Ticks represent samples from the experimental groups and their sex.

Discrimination between resistant and susceptible strains was accomplished *via* cross-validated PLS-DA modelling. The optimal model was determined to be a single-variate model with 100% accuracy (Appendix 11 for further metrics). The density plot (Figure 5.4-5-B) shows a good discrimination between resistant and susceptible strains. A subpopulation in both resistant and susceptible strains can be observed. These subpopulations are due to sex differences in adult *Ae. aegypti* species (Section 3.6). In order to probe these differences further and obtain metabolite level information, VIP scores were analysed for metabolite selection.

5.4.3.2 Key metabolites

VIP scores were calculated for all 513 bins used in the PLS-DA model. A passing threshold of 1 was applied in order to identify the most influential bins from the PLS-DA model. From the 513 bins, a total of 197 bins containing both identified and unidentified bins, scored above the threshold. Amongst the 197 bins, only 51 were identified (9.96% of whole dataset) and are shown in Figure 5.4-6. VIP selected identified bins were attributed to a total of 14

metabolites. In order to select metabolites from the model, bins with the highest representative properties were determined *via* CRS.

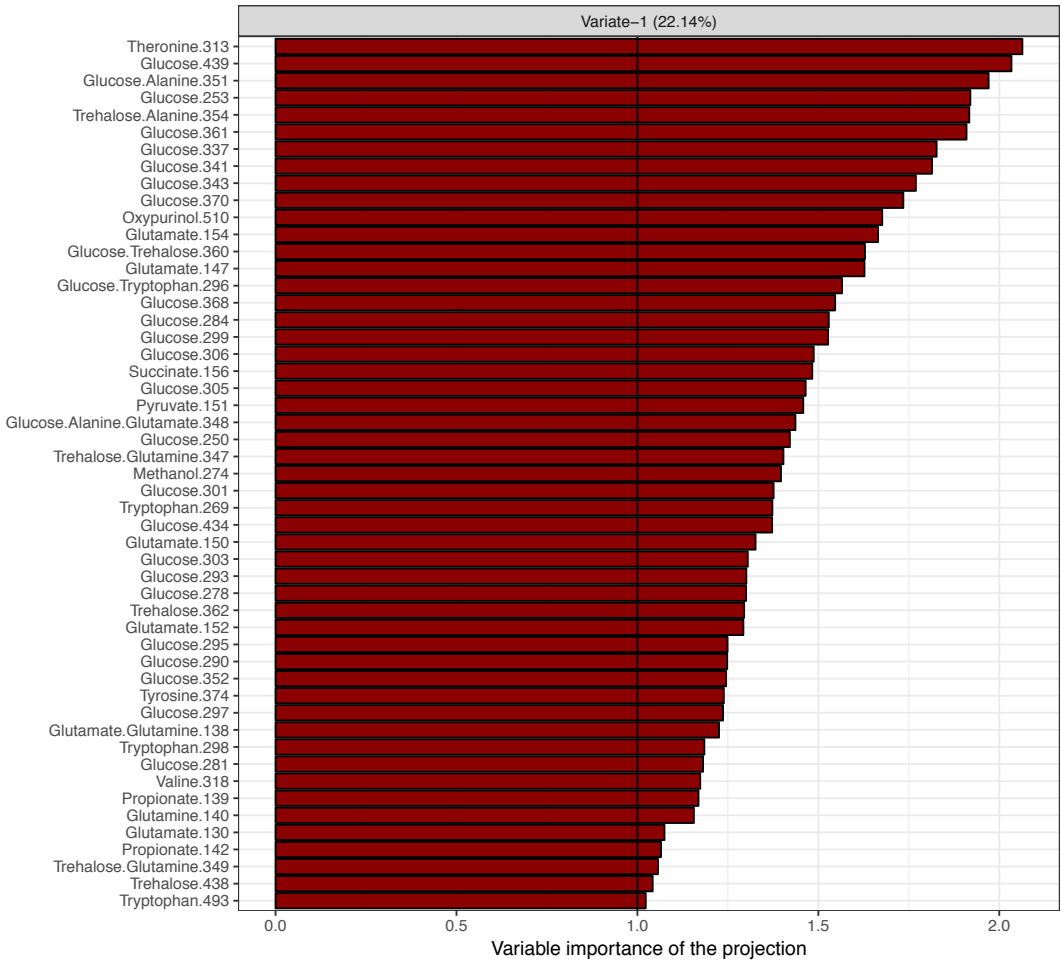


Figure 5.4-6: PLS-DA VIP scores of wild type *Ae. aegypti* adult bins. Only identified bins scoring higher than one are shown. A total of 51 identified bins out of 513 bins (9.94%) scored higher than 1, representing 14 metabolites. The black line represents the VIP score of 1.

In order to assess the representative properties of each bin corresponding to their metabolites, CRS were calculated for all identified bins. Using all CRS calculated, a passing threshold of 41.68% was determined. Only bins scoring higher than this threshold were considered as potential representatives for selected metabolites (Table 5.4-4). To create the metabolite shortlist, only non-overlapping bins (where applicable) with the highest CRS were selected.

Table 5.4-4: CRS of bins selected from *via* VIP scores for adult *Ae. aegypti*. Only the highest scoring, non-overlapping bins (where applicable) over the CRS threshold of 41.68% were considered as representative bins. Rep: representative bin.

Metabolite	Bin	CRS [%]	CRS > 41.68%	Rep	Metabolite	Bin	CRS [%]	CRS > 41.68%	Rep
Alanine	351*	72.68	✓	351	Glutamate	150	33.86	×	-
	354*	70.72	✓			147	33.69	×	
	348*	70.52	✓			152	32.90	×	
Glucose	360*	79.83	✓	368		154	31.45	×	
	296*	79.40	✓			130	30.09	×	
	368	79.09	✓			348*	6.11	×	
	278	78.92	✓			138*	3.04	×	
	305	78.82	✓		Glutamine	138*	28.64	×	-
	301	78.79	✓			140	23.93	×	
	299	78.61	✓			347*	17.44	×	
	306	78.44	✓			349*	15.96	×	
	434	78.44	✓		Methanol	274	Singlet	NA	274
	370	78.20	✓		Oxypurinol	510	Singlet	NA	510
	295	78.05	✓		Propionate	139	66.64	✓	139
	303	78.00	✓			142	43.57	✓	
	293	77.97	✓		Pyruvate	151	Singlet	NA	151
	284	77.54	✓		Succinate	156	Singlet	NA	156
	343	77.45	✓		Threonine	313	38.86	×	-
	439	77.10	✓		Trehalose	362	61.51	✓	362
	297	76.57	✓			360*	49.73	✓	
	341	76.45	✓			438	47.62	✓	
	281	75.74	✓			349*	44.36	✓	
	253	74.02	✓			347*	39.49	×	
	361	73.56	✓			354*	5.44	×	
	337	71.71	✓		Tryptophan	493	60.87	✓	493
	352	69.21	✓			298	39.68	×	
	250	67.47	✓			296*	27.40	×	
	351*	66.73	✓			269	11.23	×	
	290	61.42	✓		Tyrosine	374	69.59	✓	374
	348*	41.86	✓		Valine	318	61.64	✓	318

*: Overlapping bin

Prior to the application of the CRS passing threshold, 15 metabolites were selected *via* VIP scores. Following CRS, the metabolites glutamate, glutamine and threonine were excluded from the metabolite shortlist. Table 5.4-5 shows the metabolites selected from the PLS-DA model discriminating between resistant and susceptible strains yielding 11 metabolites from classes of alcohols, amino acids, carboxylic acids, purines, and saccharides.

Table 5.4-5: Metabolite shortlist for wild type adult *Ae. aegypti* discrimination between susceptible and resistant strains.

Class	Metabolite	Representative bin	Chemical shift [ppm]	KEGG code
Alcohols	Methanol	274	3.36	C00132
Amino acids	Alanine	351*	3.78	C00041
	Tryptophan	493	7.55	C00078
	Tyrosine	374	3.94	C00082
	Valine	318	3.16	C00183
Carboxylic acids	Propionate	139	2.17	C00163
	Pyruvate	151	2.37	C00022
	Succinate	156	2.41	C00042
Purines	Oxypurinol	510	8.27	C07599
Saccharides	Glucose	368	3.89	C00031
	Trehalose	362	3.85	C01083

*, denotes overlapping bin.

In order to demonstrate the variation represented by the selected metabolites, PCA was performed on only the bins representing the selected metabolites. Differences attributed to species can be seen in the PCA scores plot (Figure 5.4-7-A) along PC1 (55.46%). A total of 7 components were required in order to explain the 95% variance in the data. Along PC1, overlap can be observed between resistant and susceptible strains. Nevertheless, the clustering exhibited by both groups suggests through supervised modelling methods, discrimination between resistant and susceptible species can be achieved.

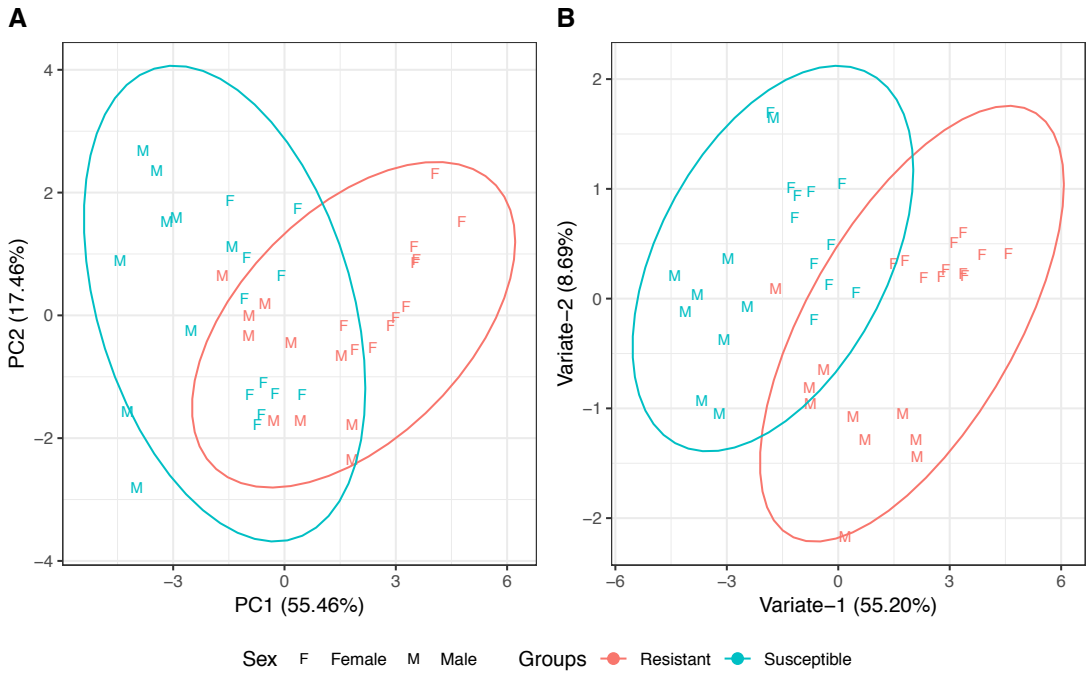


Figure 5.4-7: A) PCA scores of PC1 (55.46%) against PC2 (17.46%) for adult wild type *Ae. aegypti* ($n_{Resistant}$=21, $n_{susceptible}$=20). A total of seven components were required to explain the 95% variance in the data. Ellipses represent the 95% confidence region. B) PLS-DA scores plot of variate-1 and variate-2 discriminating between

resistant and susceptible strains of adult *Ae*. *aegypti* ($n_{Resistant}$=21, $n_{Susceptible}$=20). Two-variate model complexity was determined using cross-validation with 91.67% accuracy. Brackets report variate's explained variance. Ellipses represent 95% confidence region.

PLS-DA (Figure 5.4-7-B) on the selected metabolites was built with cross-validation revealed a two-variate model as the optimum PLS-DA model complexity for the selected metabolite data with an accuracy of 91.67% (Appendix 11 for further metrics). The discrimination of resistant species against susceptible species can be observed along a diagonal of variate-1 and variate-2.

Following the demonstration of the selected metabolites' performance on discriminating between resistant and susceptible strains, metabolite levels were compared in order to gain metabolite level information. Boxplots (Figure 5.4-8) were used to visualise metabolite levels, while BH-adjusted t-tests were used for comparison (see Appendix 22 for detailed test statistics). Metabolite level comparisons showed 11 metabolites to be significantly different between resistant and susceptible strains. More specifically, in alcohols, methanol was found to be lower in the resistant group. Of the amino acids, alanine, tryptophan, tyrosine and valine were also higher in the resistant strains. For the carboxylic acids, propionate, pyruvate and succinate were found to be higher in the susceptible group. The only purine selected was oxypurinol, which was found to be lower in the resistant group. Finally, for saccharides, both glucose and trehalose were found to be higher in the resistant group.

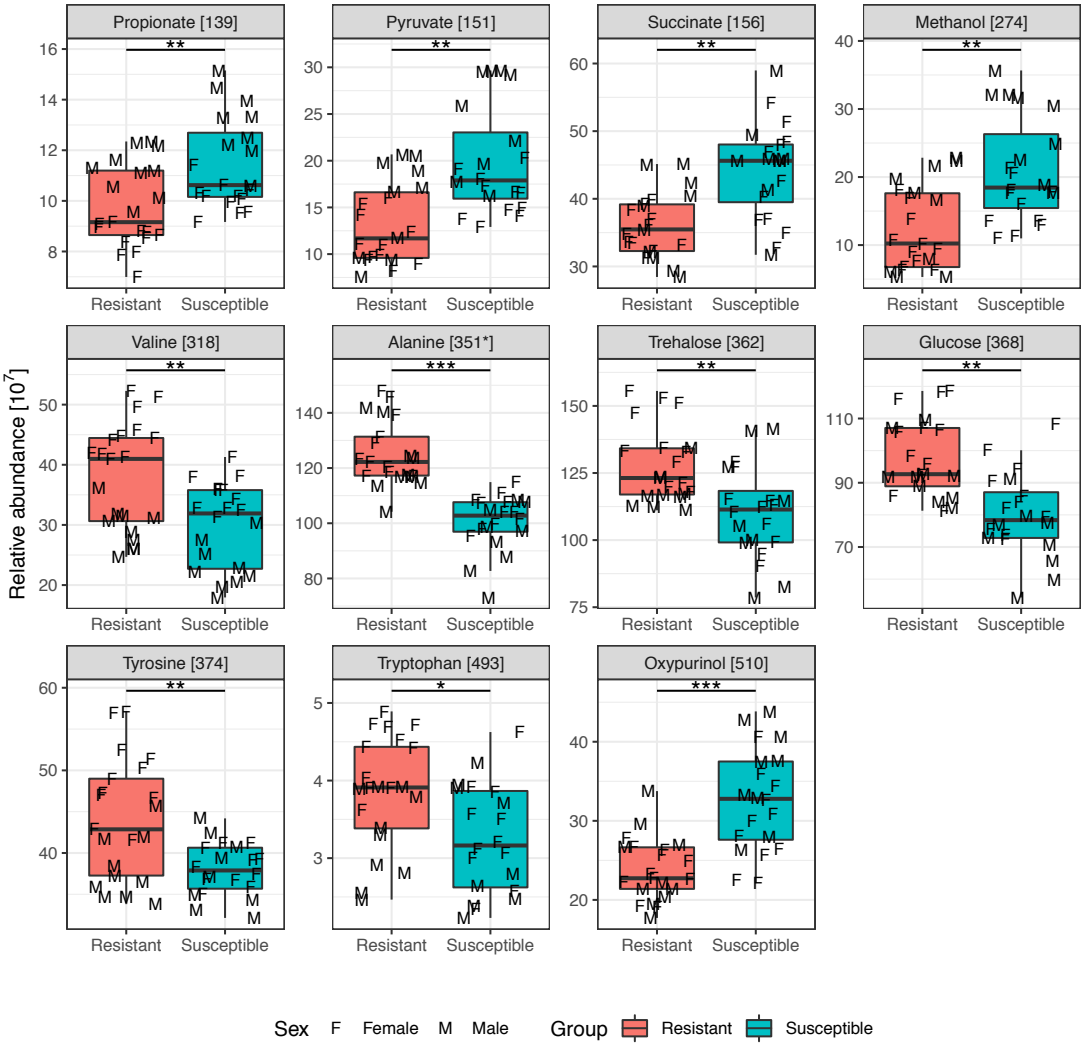


Figure 5.4-8: Boxplots for selected metabolites from the adult wild type *Ae. aegypti* ($n_{Resistant}$=21, $n_{Susceptible}$=20). *, ** and *** denotes p-values less than 0.05, 0.01 and 0.0001 respectively. * by the metabolite name denotes overlapping bin.

5.4.4 Metabolite set enrichment analysis

Selected metabolites from pupal and adult models were tabulated (Table 5.4-6) to summarise significant changes between resistant and susceptible strains. When the resistant strain was compared to the susceptible strain, the following metabolites exhibited similar changes across phases: methanol, pyruvate and oxypurinol were lower; alanine, tyrosine, glucose and trehalose were higher. All selected metabolites were used to perform MSEA.

Table 5.4-6: Summary of significant metabolite level changes between resistant and susceptible strains across pupa and adult phases. Arrows represent significant changes, and square brackets represent BH-adjusted p-values.

	Pupa		Adult	
Metabolite class	**Resistant compared to susceptible**			
Alcohols	Methanol	↓ [1.09×10^{-7}]	↓ [6.65×10^{-4}]	Methanol
Amino acids	Alanine	↑ [4.70×10^{-10}]	↑ [1.30×10^{-7}]	Alanine
	Glycine	↑ [1.87×10^{-8}]		
			↑ [1.12×10^{-2}]	Tryptophan
	Tyrosine	↑ [5.22×10^{-5}]	↑ [2.67×10^{-3}]	Tyrosine
			↑ [3.80×10^{-3}]	Valine
Carboxylic acids	Lactate	↑ [5.11×10^{-6}]		
			↓ [3.82×10^{-3}]	Propionate
	Pyruvate	↓ [4.66×10^{-8}]	↓ [4.10×10^{-4}]	Pyruvate
			↓ [4.10×10^{-4}]	Succinate
Purines	Oxypurinol	↓ [5.11×10^{-6}]	↓ [4.20×10^{-5}]	Oxypurinol
	Xanthine	↑ [2.08×10^{-7}]		
Saccharides	Glucose	↑ [4.63×10^{-12}]	↑ [1.77×10^{-4}]	Glucose
	Trehalose	↑ [5.22×10^{-5}]	↑ [1.71×10^{-3}]	Trehalose

MSEA (Figure 5.4-9) was performed with EASE correction and BH adjustment for the p-values in order to generate a list of pathways which represent the differences between resistant and susceptible species on a metabolic pathway level (Table 5.4-7). A total of 11 pathways were over-represented between adult and pupae, of which three were unique to pupae, four were unique to adults, and four were common.

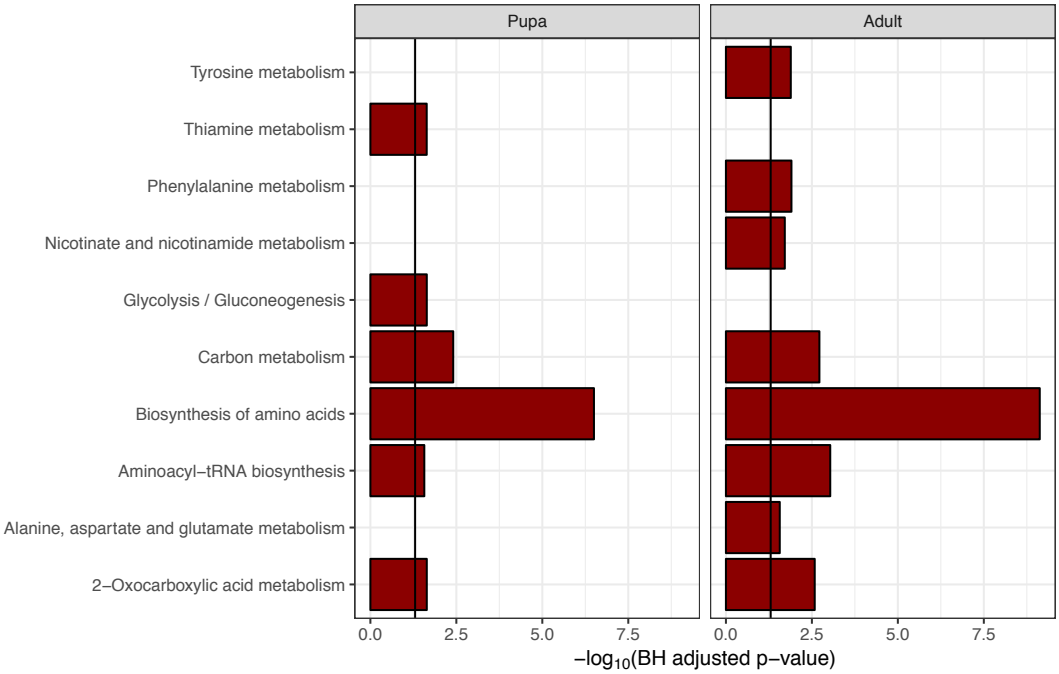


Figure 5.4-9: Selected metabolites analysed *via* MSEA from both pupal and adult phases. Resulting p-values were adjusted with BH for multiple testing. Black line represents p-value 0.05.

Table 5.4-7: Details of the MSEA for wild type *Ae. aegypti* reporting stage, raw & BH adjusted p-values for Fisher's exact test, number of metabolite hits per pathway and matching metabolites.

Pathway	Stage	Raw p-value	BH adjusted p-value	Hit/total (%)	Matches
2-Oxocarboxylic acid metabolism	Pupa	2.3×10^{-3}	2.3×10^{-2}	2/134 (1.49%)	Pyruvate, Tyrosine
	Adult	2.8×10^{-4}	2.6×10^{-3}	4/134 (2.99%)	Pyruvate, Tryptophan, Tyrosine, Valine
Alanine, aspartate and glutamate metabolism	Adult	5.7×10^{-3}	2.7×10^{-2}	3/29 (10.34%)	Pyruvate, Alanine, Succinate
Aminoacyl-tRNA biosynthesis	Pupa	4.1×10^{-3}	2.7×10^{-2}	3/52 (5.77%)	Glycine, Alanine, Tyrosine
	Adult	4.9×10^{-5}	9.3×10^{-4}	4/52 (7.69%)	Alanine, Tryptophan, Tyrosine, Valine
Biosynthesis of amino acids	Pupa	7.9×10^{-9}	3.2×10^{-7}	4/128 (3.12%)	Pyruvate, Glycine, Alanine, Tyrosine
	Adult	2.0×10^{-11}	7.5×10^{-10}	5/128 (3.91%)	Pyruvate, Alanine, Tryptophan, Tyrosine, Valine
Carbon metabolism	Pupa	1.9×10^{-4}	3.9×10^{-3}	4/112 (3.57%)	Pyruvate, Glycine, Alanine, Methanol
	Adult	1.5×10^{-4}	1.9×10^{-3}	4/112 (3.57%)	Pyruvate, Alanine, Succinate, Methanol
Glycolysis / Gluconeogenesis	Pupa	2.3×10^{-3}	2.3×10^{-2}	3/31 (9.68%)	Pyruvate, Glucose, Lactate
Nicotinate and nicotinamide metabolism	Adult	3.6×10^{-3}	1.9×10^{-2}	3/55 (5.45%)	Pyruvate, Succinate, Propanoate
Phenylalanine metabolism	Adult	1.6×10^{-3}	1.2×10^{-2}	3/74 (4.05%)	Pyruvate, Succinate, Tyrosine
Thiamine metabolism	Pupa	2.9×10^{-3}	2.3×10^{-2}	3/32 (9.38%)	Pyruvate, Glycine, Tyrosine
Tyrosine metabolism	Adult	2.1×10^{-3}	1.3×10^{-2}	3/79 (3.8%)	Pyruvate, Succinate, Tyrosine

5.5 Chapter results summary

In this chapter, wild type resistant and susceptible strains of *An. gambiae* and *Ae. aegypti* species were compared using NMR metabolomics. Overall, wild type PLS-DA models demonstrated higher discriminatory properties (Table 5.5-1) between resistant and susceptible strains when wild type models. All models were found to be robust, exhibiting high accuracy (highest: *Ae. aegypti* pupae and adult full dataset with 100%, lowest: *Ae. aegypti* adult selected with 64.29%). The *An. gambiae* pupal model, built with the full dataset, yielded a list of 16 selected metabolites compared to the adult model where only nine were selected. In the *Ae. aegypti* full dataset models, pupae selected 10 and adults selected 11 metabolites.

Table 5.5-1: Summary of cross-validated PLS-DA model performance and following metabolite selections. VIP selected bins show the number of identified bins selected by VIP scoring. Brackets show the percentage of identified bins amongst all VIP selected bins. NA: not applicable.

Species	Phase	Data coverage	Model variates	Accuracy	VIP selected bins	Number of metabolites selected
An. gambiae	Pupa	Full dataset	1	64.29%	52 (28.57%)	16
	Adult		2	87.50%	30 (19.11%)	9
	Pupa	Selected	1	85.71%	NA	NA
	Adult	metabolites	2	93.75%	NA	NA
Ae. aegypti	Pupa	Full dataset	1	100.00%	60 (24.49%)	10
	Adult		1	100.00%	51 (25.89%)	11
	Pupa	Selected	1	87.50%	NA	NA
	Adult	metabolites	2	91.67%	NA	NA

Overall, the metabolite selection performed on pupal models yielded more metabolites compared to the adult models. This indicates that in both species the differences between resistant and susceptible strains were more pronounced in pupae.

Interestingly, the selected metabolites which are common between both strains of *An. gambiae* and *Ae. aegypti*, all show a reverse trend between species (Table 5.5-2). Uniquely significant metabolites were only identified in pupae of *An. gambiae*. These were glutamate, isoleucine, threonine, acetate, formate and fumarate.

Table 5.5-2: Summary table for significant metabolite level changes represented as arrows between species of *An. gambiae* and *Ae. aegypti*. Comparisons represented as levels in resistant strains compared to susceptible strains for pupae and adults. Square brackets represent BH-adjusted p-values.

		An. gambiae		***Ae. aegypti***	
		Resistant : Susceptible			
Class	**Metabolite**	**Pupa**	**Adult**	**Pupa**	**Adult**
Alcohols	Methanol	↑ [8.17×10^{-9}]	↓ [1.90×10^{-4}]	↓ [1.09×10^{-7}]	↓ [6.65×10^{-4}]
Amino acids	Alanine	↓ [2.23×10^{-4}]		↑ [4.70×10^{-10}]	↑ [1.30×10^{-7}]
	Glutamate	↑ [7.22×10^{-4}]			
	Glycine	↓ [1.25×10^{-3}]		↑ [1.87×10^{-8}]	
	Isoleucine	↓ [2.89×10^{-3}]			
	Threonine	↓ [1.48×10^{-4}]	↓ [1.67×10^{-3}]		
	Tryptophan	↑ [1.67×10^{-3}]			↑ [1.12×10^{-2}]
	Tyrosine		↓ [4.28×10^{-2}]	↑ [5.22×10^{-5}]	↑ [2.67×10^{-3}]
	Valine				↑ [3.80×10^{-3}]
Carboxylic acids	Acetate	↑ [7.73×10^{-3}]			
	Formate	↑ [2.72×10^{-3}]			
	Fumarate	↑ [4.05×10^{-8}]			
	Lactate	↓ [6.12×10^{-5}]	↓ [1.07×10^{-2}]	↑ [5.11×10^{-6}]	
	Propionate	↑ [9.64×10^{-3}]	↑ [9.19×10^{-3}]		↓ [3.82×10^{-3}]
	Pyruvate	↓ [3.79×10^{-10}]	↓ [9.50×10^{-6}]	↓ [4.66×10^{-8}]	↓ [4.10×10^{-4}]
	Succinate		↑ [1.03×10^{-3}]		↓ [4.10×10^{-4}]
Purines	Oxypurinol	↑ [1.19×10^{-3}]	↑ [1.10×10^{-4}]	↓ [5.11×10^{-6}]	↓ [4.20×10^{-5}]
	Xanthine			↑ [2.08×10^{-7}]	
Sugars	Glucose	↓ [1.48×10^{-4}]		↑ [4.63×10^{-12}]	↑ [1.77×10^{-4}]
	Trehalose	↑ [1.95×10^{-3}]	↑ [9.19×10^{-3}]	↑ [5.22×10^{-5}]	↑ [1.71×10^{-3}]

A summary of MSEA analyses are tabulated in Table 5.5-3. The following pathways were found to be commonly over-represented in all the groups studied: carbon metabolism, biosynthesis of amino acids, aminoacyl-tRNA biosynthesis, and 2-oxocarboxylic acids metabolism. In *An. gambiae*, pupae-specific pathways are: 2-Oxocarboxylic acid metabolism, ABC transporters, alanine, aspartate & glutamate metabolism, aminoacyl-tRNA biosynthesis, butanoate metabolism, FoxO signalling pathway, glycine, serine & threonine metabolism, glycolysis / gluconeogenesis, glyoxylate & dicarboxylate metabolism, neuroactive ligand-receptor interaction, nitrogen metabolism, phosphonate & phosphinate metabolism, taurine & hypotaurine metabolism, and valine, leucine & isoleucine biosynthesis. In comparison, only phenylalanine metabolism was unique to adults. Meanwhile, nicotinate & nicotinamide metabolism, propanoate metabolism, and pyruvate metabolism were common for both pupae and adults. In *Ae. aegypti*, common pathways between pupae and adults are: 2-Oxocarboxylic acid metabolism and aminoacyl-tRNA biosynthesis. Pupae specific pathways of *Ae. aegypti* are: glycolysis / gluconeogenesis, glycine, serine & threonine metabolism and thiamine metabolism. Adult specific pathways are: alanine, aspartate & glutamate metabolism, nicotinate and nicotinamide metabolism, phenylalanine metabolism and tyrosine metabolism.

Table 5.5-3: Summary of MSEA analyses showing pathways identified differing between resistant and susceptible strains of *An. gambiae* and *Ae. aegypti*.

Pathway	***An. gambiae***	***Ae. aegypti***
2-Oxocarboxylic acid metabolism	P	C
ABC transporters	P	
Alanine, aspartate and glutamate metabolism	P	A
Aminoacyl-tRNA biosynthesis	P	C
Biosynthesis of amino acids	C	C
Butanoate metabolism	P	
Carbon metabolism	C	C
FoxO signalling pathway	P	
Glycine, serine and threonine metabolism	P	P
Glycolysis / Gluconeogenesis	P	P
Glyoxylate and dicarboxylate metabolism	P	
Neuroactive ligand-receptor interaction	P	
Nicotinate and nicotinamide metabolism	C	A
Nitrogen metabolism	P	
Phenylalanine metabolism	A	A
Phosphonate and phosphinate metabolism	P	
Porphyrin and chlorophyll metabolism	P	
Propanoate metabolism	C	
Pyruvate metabolism	C	
Taurine and hypotaurine metabolism	P	
Thiamine metabolism		P
Tyrosine metabolism	A	A
Valine, leucine and isoleucine biosynthesis	P	

A: only in adults, C: common for both pupae and adults, P only in pupae.

5.6 Chapter discussion

Employing an NMR metabolomics approach, metabolite profiles of the wild type *An. gambiae* and *Ae. aegypti* were observed for early pupae and early adults. Surprisingly, there is no metabolic profiling of resistant and susceptible strains of wild type mosquitoes to date. Current publications on mosquito metabolomics usually target a specific case, such as blood digestion [164] or pathogen infection [149]. Using an NMR metabolomics approach to explore the metabolite profiles of wild type resistant and susceptible strains, differences between distinct profiles and a potential resistance biomarker was proposed.

This study was based on the assumption that the different resistance statuses would manifest different metabolic signatures. Identified pathways characteristic to resistance status were correlated to pathways demonstrated by KD16 and KD17 datasets. Each dataset presented metabolic profile differences observable *via* NMR on a single pupa/mosquito scale. Between the pupa and adult datasets, metabolic profile separation through PCA attributable to resistance status was more apparent in pupae compared to adults. The VK7 strain of *An. gambiae* has been shown to have cuticular resistance as well as metabolic and target site resistance *via* several studies [213], [214]. If only MSEA results are considered, common identified pathways between the KD strains and the wild type strains of *An. gambiae*, supports the presence or these resistance types. Although, it should be noted that this link cannot be directly established through metabolite levels since the metabolite levels of a knock-down state would not be representative of a wild type profile. Nevertheless, both in knock-down and wild type resistance states, similar pathways would be active and over-represented.

In this chapter, metabolic profiles of resistant and susceptible wild type *An. gambiae* and *Ae. aegypti* were compared during pupal and adult stages. For both pupal and adult comparisons, the metabolic differences between resistant and susceptible strains were observable through NMR metabolomics. In both *An. gambiae* and *Ae. aegypti* datasets, metabolic profile differences attributable to resistance status in PCA scores were more noticeable in pupae than in adults. In *Ae. aegypti*, cluster separation attributable to resistance status corresponded to 48.21% explained cumulative variance in pupae, while in adults this was 42.31%. Such high explained variation indicates a robust PLS-DA model with high accuracy. All data wild type models scored accuracy ranging from 64.29%-100%, whereas models built on selected metabolites scored a narrower range of 85.71%-93.75%

showing the selection of metabolites captured the differences between resistant and susceptible strains. Using the metabolites selected from the PLS-DA models of pupae and adults, MSEA was performed in order to gain insight on the differences between resistant and susceptible strains at a pathway level. Common pathways over-represented between pupae and adults were: carbon metabolism, biosynthesis of amino acids, aminoacyl-tRNA biosynthesis, and 2-oxocarboxylic acids metabolism. In Chapter 4, the same set of pathways were identified in the comparison between knock-down and control strains. These pathways were attributed to the production of CHC in the knock-down *An. gambiae* (Section 4.6). Identification of these pathways in the MSEA comparison of resistant and susceptible strains of wild type *An. gambiae* might suggest the presence of CHC resistance in *An. gambiae* and in the VK7 strain of *Ae. aegypti,* which was suggested through transcriptomics studies [213], [214].

Compared to the previous section in this thesis wild type *An. gambiae* and *Ae. aegypti* showed great agreement between the pupae and adults of the same species. In *An. gambiae* threonine and lactate was significantly lower in resistant pupae and adults. Meanwhile propionate, oxypurinol, and trehalose was significantly higher in resistant pupae and adults. Surprisingly in *Ae. aegypti* the same list of agreeing metabolites is quite different with methanol, pyruvate, and oxypurinol is significantly lower while alanine, tyrosine, glucose, and trehalose is significantly higher. Between *An. gambiae* and *Ae. aegypti* xanthine levels are significantly higher and lower respectively. Although, MSEA result did not show any pathways related xanthine. It is possible that metabolites found together with xanthine might be some of the unidentified metabolites in the dataset. Interestingly, regardless of the dissimilar list of metabolites, MSEA resulted with pathways that are over-represented between resistant and susceptible strains of both species. Glycolysis/gluconeogenesis and its vicinity pathways glycine, serine and threonine metabolism; and 2-oxocarboxylic acid degradation were significantly over-represented between resistant and susceptible strains of pupal *An. gambiae* and *Ae. aegypti*. For adults MSEA results were more clustered around amino acids. The pathways over-represented in adults were tyrosine metabolism, amino acid biosynthesis, and phenylalanine metabolism. Although, the metabolite list may not have high agreement between the species, there are clear structure in them. This kind of data is typically amenable to statistical modelling. Unfortunately, these two datasets do not have identical set of bins. Thus, creating a merged dataset with proper verification of would be better suited for future work. Nevertheless, the data suggest that between resistant

susceptible strains of these species, trehalose an important metabolite for energy generation and protection against desiccation is the only constantly changing metabolite.

Trehalose being significantly higher in resistant strains (pupae and adult) of *An. gambiae* and *Ae. aegypti*, which may be instrumental in identifying resistant and susceptible species in the wild type populations. In *An. gambiae*, the mean relative intensities recorded for the selected trehalose bins ranged from 2.98 x 10^8 to 3.59 x 10^8 in pupae and in adults from 6.21 x 10^8 to 8.68 x 10^8 between susceptible and resistant strains. Values recorded for the *Ae. gambiae* strains were from 8.99 x 10^8 to 10.3 x 10^8 in pupae and from 11.0 x 10^8 to 12.8 x 10^8 in adults. When univariate tests were performed, these mean values were found to be significantly higher in resistant species, even after multiple test adjustments. This shows that regardless of how close these values might be, the variation within the strains is narrow enough to be statistically significant, thus indicating trehalose as a putative biomarker for identification of resistant strains. It should be noted that, although the mean values of trehalose are always higher in resistant strains, there isn't a consistency between the range they are higher. This means that a classifier threshold of 3.25 x 10^8 might yield a high accuracy in identification of resistant *An. gambiae* pupae, but the very same value is most likely to class all the adult strains as resistant. In order to avoid such cases, one or two or even a combination of preventative steps should be taken. These steps could include the use of specific threshold levels unique to different species/strains/stages or normalisation of the trehalose measurement to another metabolite. Using different thresholds for identification would require more testing and the establishment of this level for a variety of different cases, although once established this would make the measurements simpler as it would involve the measuring of one metabolite instead of multiple. On the other hand, using another metabolite for normalisation might make the optimal resistance identification threshold more applicable to a variety of species and/or stages. Although using a single value for identification is practically easier, more than one metabolite needs to be measured which creates an additional step to the method. Furthermore, such a metabolite needs to be similar to the trehalose profile in the species/stage it's being normalised in. Hence, any candidate metabolite for normalisation ideally should follow the same trend across all strains/stage (e.g. all non-significant or all significantly higher/lower but not alternating levels) and have similar variation.

Significantly high levels of trehalose may play a critical role in resistant strains from wild type populations. Numerous studies [215]–[217] have linked trehalose to a role in desiccation and

stress response. As trehalose is the primary sugar in the haemolymph and the main source of energy, it is kept at a constant level [210]. When in excess, trehalose is converted to glycogen for storage [210]. For this balance to be maintained, trehalose needs to be efficiently transported between cells and the haemolymph. This is performed by members of the ABC transporter family. From the significantly over-represented MSEA pathways between resistant and susceptible strains only one result included trehalose, which is the ABC transporters. In this family, trehalose can be transported *via* four possible transporters TreS, TreT, TreU and TreV. Amongst these, TreT type transporters are commonly observed in insects. TreT-1 was previously observed in *Drosophila melanogaster* [218] and in *An. gambiae* [219]. In *Ae. aegypti*, TreT-1 was inferred from uptake assays [220]. TreT-1 is transmembrane transporter and was shown to be activated depending on the H^+ levels in a cell or haemolymph [221]. It should be noted that this transporter family was only identified for the *An. gambiae* pupae. This is most likely due to the fact that the ABC transporter family is a collection of metabolites and transporters. Selected metabolites for the other datasets are not comprised of enough metabolites that are members of the ABC transporter 'pathways' to yield a significant BH adjusted p-value.

Using Tret-1 as an insecticide target may be an efficient approach to develop the next generation of insecticides. Targeting Tret-1 is a favourable approach due to its importance in mosquitoes and its relative unimportance in vertebrates. Hence, insecticides targeting Tret-1 potentially could be less or near negligibly toxic for vertebrates. Although, given the importance of trehalose in mosquitoes [222], downstream metabolism of trehalose may be a more suitable target. After trehalose is exported to the haemolymph, it is broken down in to glucose which is then taken up by the cell for energy. Metabolism of trehalose is facilitated by one of two trehalase enzymes Tre-1 and Tre-2 [222]. Both of these enzymes hydrolyse trehalose into α-D-glucose and a β-D-glucose [222]. The main differences between these two enzymes is Tre-1 is the soluble form where it has been identified in haemolymph [223], and Tre-2 is a membrane-bound enzyme where it is found on the extracellular side of trehalose utilising cells such as flight muscle cells [210]. It has been hypothesised that trehalase activity is at least partly regulated by hormones such as the set of hormones collectively called 'juvenile hormones' [224].

An advantage of using trehalose-related insecticides is that trehalose does not play as critical a role in vertebrates as it does in insects, hence overcoming toxicity in mammals [225]. However, trehalose has been shown to play important roles in plants as well (e.g. protection

against desiccation [226] and increased yield [227]), although mechanisms are not fully characterised to date. Validoxylamine [228] is an efficient inhibitor of trehalases [229] but finding delivery methods for it have proven to be challenging. Due to its polar nature, delivery of the inhibitor through the CHC layer made the drug inefficient in applications. Alternatively, shifting the focus from mosquitoes to plant-feeding pests; trehalose-related insecticides can be delivered systemically. In this way, the drug would be delivered through ingestion, where the polar nature of the compound becomes an advantage.

Chapter 6

6 General Conclusions

The main objective of this investigation was to employ nuclear magnetic resonance (NMR) metabolomics to advance the field of mosquito metabolomics by increasing understanding of insecticide resistance. In order to achieve this, I first established a robust method to consistently monitor and compare metabolic profiles in individual mosquitoes (pupae and adult). I subsequently compared metabolite profiles of knock-down and control *An. gambiae* probing the cuticular hydrocarbon (CHC) production which plays a critical role in insecticide penetration and delivery. Metabolic profiles of wild type *An. gambiae* and *Ae. aegypti* with pyrethroid resistance were then compared to explore whether conserved pathways were affected.

Prior to this work, there was very little known about the metabolite profiles of CHC biosynthesis and pyrethroid resistant species (*An. gambiae* and *Ae. aegypti*). A great deal of work has been undertaken in various insects in order to characterise the mechanism of CHC biosynthesis (covered extensively in [78]). Alterations in the cuticular layer of mosquitoes is an area that has been gaining more attention due its link to insecticide resistance [74]. Mosquito-specific cytochrome P450s, Cyp4g16 and Cyp4g17, have recently been shown to play a critical role in CHC biosynthesis pathway [75]. The link between CHC biosynthesis and the metabolome has not been studied to date. The need to understand this mechanism in greater detail, as well as providing an explanation for other potential (nuanced) contributors to insecticide resistance, led to the natural progression of this study to a global approach on individual insects at different life stages *via* NMR metabolomics.

6.1 Summary of thesis findings

6.1.1 Sex-specific differences vary amongst the mosquito species

As a part of the data exploration in this project sex-specific differences in mosquitoes were investigated. Unexpectedly, these differences were attenuated in the knock-down groups. This was interpreted as either an effect of the Gal4/UAS system or an indirect effect on fatty acid biosynthesis. On the contrary, in wild type *Ae. aegypti* these differences were accentuated. Unfortunately, the experimental design for this study is not capable to answer this question. Furthermore, four metabolites were found to be consistently significantly

different between males and females across all comparisons. These metabolites acetate (higher in males), lactate (higher in females), propionate (higher in males), and glucose (higher in females). These metabolites are mostly known for their roles in energy generation and storage. High levels of glucose in females was interpreted as higher potential of both energy generation (through glycolysis) and storage (through generation of glycogen). Similarly, high lactate is a typical indicator of anaerobic respiration and is converted from acetate thus high energy demand. Propanoate metabolism synthesis a wide range of fatty acid precursor, thus fluctuations in propionate can be speculated to be an indicator of HCs synthesis differences, although this would require a different experimental setup to test.

6.1.2 Cyp4g16 and Cyp4g17 are temporally active decarbonylases with potential specificity for branched hydrocarbons

Supporting evidence of decarbonylase function of Cyp4g16 and Cyp4g17 was shown through a GAL4/UAS knock-down model in *An. gambiae.* By monitoring the precursors metabolites of CHC biosynthesis valine, leucine, isoleucine, and acetate activities of Cyp4g16 and Cyp4g17 were observed. By impairing the CHC production accumulation of precursor metabolites were expected. Amongst the precursors, sole accumulation of valine (a branched amino acid), showed evidence of both Cyp4g16's (experimentally shown [75]) and Cyp4g17's (hypothesised) activity as decarbonylases. Furthermore, through the investigation of these knock-down strains in pupal and adult stages, their temporal activity was observed. Higher valine levels were observed in Cyp4g16 knock-downs of pupae and adults supporting evidence for Cyp4g16's higher activity in pupae [75]. Similarly, an accumulation of valine was observed only in the Cyp4g17 adults proposing a higher activity in adults.

6.1.3 Pyrethroid resistant mosquito species have distinct metabolic profiles compared to susceptible species

Lastly, to explore the polar metabolic profiles of resistant strains of *An. gambiae* and *Ae. aegypti* were investigated. This novel work aimed to give insight on the neglected polar metabolic aspect of resistance. By using the pipeline developed in this project. Resistant and susceptible strains of *An. gambiae* and *Ae. aegypti* were studied as pupae and adults. Both species showed great performance in statistical model building allowing for effective metabolite selection. It was expected to see relatively similar metabolic profiles between two species due to the consistent growth conditions and being bred in the laboratory for many years. Although, this was not the case. Interestingly the metabolic profiles between different

stages of the same species were very close. This was not observed in the previous comparisons (Sex or Cyp4g knock-down). Upon performing MSEA a set of pathways similar to Cyp4g knock-downs was obtained. It is known that *An. gambiae* resistant strain (VK7) possessed cuticular resistance. It was expected to see similar set of pathways such as amino acid biosynthesis, carbon metabolism, and valine, leucine and isoleucine biosynthesis. This observation suggested the presence of cuticular carbon resistance on the *Ae. aegypti* Cayman strain although this requires verification. Perhaps, the most interesting outcome of this section is trehalose. This metabolite was found significantly higher across all comparisons and were selected by all statistical models. This study proposes trehalose as a potential biomarker for identification of resistance.

6.2 Critical evaluation of methods

6.2.1 Metabolomics and metabolite coverage

For any study, limitations should be acknowledged. It should be noted that no single analytical method can capture the entirety of a metabolic profile [230]. Metabolite coverage is one of the main limitations as there are multiple extraction methods widely used in metabolomics. For polar metabolites the most common extraction methods are: methanol:water (4:1, v:v) [231], acetonitrile:water (1:1, v:v) [189] and chloroform:methanol:water (1:1:1, v:v:v) [232]. In this study, the observable metabolites were polar metabolites extracted with 50% acetonitrile in water. Any metabolite requiring a different extraction procedure would not be detected, such as non-polar compounds. Chloroform:methanol:water is typically a two stage extraction method which has the added benefit of extracting both polar and non-polar metabolites although this method requires aliquoting from two separate layers without disturbing the other layer and is prone to inconsistent aliquoting. This, compounded with the fact that focus of this project was on polar metabolites, meant this extraction method was not preferred. Between the extraction methods of acetonitrile:water and methanol:water, acetonitrile:water was preferred due to its higher metabolite coverage [233]–[236]. It should be noted that acetonitrile:water can extract some mobile lipids alongside the polar metabolites. Furthermore, sonication was used for homogenisation of the samples which can heat up the samples during the homogenisation process. In order to address this problem, sonication was performed on an ice bath with 30 s breaks taken in between sonicator pulses.

This study utilised NMR as the metabolite detection method of choice. As described (Section 1.6.5.1), NMR is inherently less sensitive compared to mass spectrometry (MS). However, NMR is more reproducible and self-contained as samples do not come in contact with the detection probe which makes it virtually impossible for sample cross-contamination during data acquisition to occur. It should be noted that ideally these techniques would be used together as complementary methods. Unfortunately, for this study this was unfeasible due to time and financial constraints, as it is for most studies.

In this study, the scope of observable metabolites was limited to the coverage of acetonitrile:water extraction and observation *via* 700 MHz magnet NMR fitted with an inverse cryoprobe. This method does not capture low concentration metabolites (i.e. secondary metabolites) or non-polar metabolites. Identification was carried out using an in-house library, Chenomx and online databases such as biological magnetic resonance bank (BMRB). The in-house library has the advantage of standards being acquired on the same spectrometer and so better matching between the standards and the samples can be achieved. The main disadvantage of the library is its smaller size compared to other libraries. External databases (i.e. BMRB) cover a wide range of metabolite spectra, but these spectra are acquired under different conditions such as: pH, temperature and magnetic field which makes the matching of the spectra a challenging process. Chenomx is currently the 'gold standard' software for NMR metabolite identification. However, it is better suited for the identification of mammalian metabolites, as is the case with most metabolite identification software. Although it is more favourable to establish a mosquito metabolome library, this would be a great undertaking that would merit a PhD on its own.

6.2.2 Identification of unknowns

In any untargeted metabolomics approach, unidentified metabolites are a challenge. Unidentified metabolites can be categorised as known unknowns and unknown unknowns. In an NMR context, a known unknown is easier to deal with than an unknown unknown. A known unknown is when all peaks arising from the unknown metabolite are free from overlap and are able to be picked out making elucidation of the structure and identification of the molecule a possibility. A unknown unknown is when this is not possible since the information to elucidate the structure of the metabolite is overlapped beyond deconvolution or it is simply not possible to identify all the signals arising from the unknown metabolite. It is sometimes possible to recover some information *via* statistical approaches such as ANOVA

or t-test as one would expect a significantly different unknown bin would have all other bins associated to it to be also significantly different. Although this is true in certain cases, such as when a great percentage of bins are identified, usually it is not the case in the majority of NMR metabolomics studies and it is not the case in this one.

In this project, two distinctly different datasets were used (*An. gambiae* and *Ae. aegypti*) and within these species' datasets, only metabolite levels were fluctuating. For the *An. gambiae* dataset, only 107 out of 496 (21.57%) bins were identified, leaving 389 bins to be classed as unknown unknowns. For the *Ae. aegypti* dataset, 111 out of 517 (21.47%) bins were identified, leaving 406 unidentified. Excluding the possible overlaps, theoretically this can represent any number of compounds between 1-389 for *An. gambiae* and 1-406 for *Ae. aegypti*. Unfortunately, due to the large quantity of unidentified bins, the possible bin combinations would not be able to be deconvoluted with the significance shown by ANOVA or t-test. Furthermore, due to the multiple comparisons performed, a p-value adjustment would have to be carried out prior to making assumptions for the grouping of bins. Given the large number of comparisons, some bins, especially overlapping ones, may score higher than 0.05 while other bins from the same molecule may be scoreless or *vice versa*. Additionally, similar significant differences from different molecules may be wrongly interpreted as one metabolite. Subsequently, to be more conservative, this approach was not considered in this project.

Nevertheless, it is not impossible to identify the unknowns. In order to identify unknowns, a more natural products approach can be taken. This requires the use of various chromatographical methods to either isolate or simplify the complexity of a sample which can then be analysed using a combination of NMR and MS. Using a variety of NMR experiments (HSQC, HMBC, TOCSY, COSY, DEPT45 and DEPT135), almost all isolated compounds, or even simplified mixtures of two to five compounds, can be identified and verified *via* MS. Meanwhile, MS can identify a great deal of low concentration compounds which may not be detectable by NMR.

6.2.3 Relative quantification over absolute quantification

One of the core properties of a metabolomics study is the quantification of metabolites, be it relative or absolute. Within the most popular metabolomics platforms of NMR and MS, NMR is more amenable for quantification. Besides the obvious method of using a calibration

curve, NMR can exploit signal intensities being directly proportional to compound concentration, yielding absolute quantification without a calibration curve. This is typically done in one of two ways on ^{1}H spectra.

The first method uses an internal standard in the sample. Commonly used NMR internal standards include; trimethylsilylpropionate (TSP), trimethylsilylpropanesulfonate (DSS), and trimethylsilane (TMS). Typically, the internal standard chosen will resonate in a region where molecules of interest do not resonate. This allows the integration of the internal standard without the worry of signal overlap. Once the signals arising from the molecule of interest are assigned and integrated, both molar ratios of the internal standard and the molecule of interest can be calculated and extrapolated to absolute quantification with the additional knowledge of molecular weight and sample volume.

The second method is perhaps the most convenient method and is known as electronic reference to access *in vivo* concentrations (ERETIC). This method requires the preparation of an external standard sample with known concentration. This external sample is then measured separately from the measurement of the molecule of interest. Benefiting from the consistency of NMR measurement, the spectral information of the external standard can be used to quantify the molecule of interest. A caveat to this method is, the external standard needs to be prepared in the same way as the molecule of interest and data acquisition needs to be performed under identical conditions if possible, or near identical with the trade-off of accuracy.

These methods are simple and effective for pure or near pure compounds. However, metabolomics data is often with noise, many overlaps, and sometimes with shimming or baseline issues which can affect absolute quantification. Integration of non-overlapping samples with good integration regions is not a common occurrence in NMR metabolomics data, hence using a relative quantification can be done without the worry of searching for the perfect signal. Furthermore, relative quantification allows the use of unidentified signals and so the entire dataset can be used. When using an absolute quantification method, all unidentified metabolites need to be discarded prior to data analysis. In this project, due to the very low number of identified metabolites and high overlapping signals, relative quantification was preferred in order to be able to use the entire dataset in analyses.

6.2.4 Samples

Due to the nature of the samples, a considerable number of samples failed the QC criteria. As a result of this, pupae dataset comparison of knock-down strains were lower powered, which decreased the accuracy in the models.

One of the limitations in this project was the use of only a loss of function model rather than being able to also analyse data from a gain of function system, for example by Cyp4g overexpression. An over-expression model would have provided a system to compare and verify the results obtained with the RNAi knockdown results. However, both Cyp4g16 and Cyp4g17 enzymes are already highly expressed in the wild type susceptible mosquitoes [188], and when overexpression was examined using the available oenocyte driver (Gareth), no significant increase in Cyp4g protein was observed (personal communication with Dr Gareth Lycett). For example, in order to increase the expression level of these enzymes two-fold, a Gal4 line would be required that could drive expression at a similar level to the native Cyp4g proteins. No such driver is available though. In contrast, RNA interference (RNAi) is significantly more efficient due to the fact that a single dsRNA is capable of knocking down many transcripts. Since over-expression was not feasible, investigation of metabolic consequences on CHC biosynthesis through knocking down of Cyp4g16 and Cyp4g17 was the only option.

The wild type species used in this study were collected from different geographical locations. A caveat is that the resistance status and metabolic differences caused by different geographical backgrounds cannot be deconvoluted. In this study, it is assumed that any metabolic differences originating from geographical origins are negligible as a consequence of laboratory breeding over years (*An. gambiae* > 5 years and *Ae. aegypti* > 8 years).

6.2.5 Statistics

As a multivariate data type, metabolomics data should be analysed largely using multivariate statistics that can be complemented with univariate testing. There are examples in the literature [237] in which univariate testing is used as a way to reduce dimensionality by selecting significant variables. However, with multivariate data, there are a great number of comparisons. Subsequently, following p-value correction for multiple testing, a great deal of power is lost in the study. Furthermore, univariate methods are not designed to capture the relations between variables otherwise explained with multivariate methods.

In this thesis, a combination of multivariate transformations (i.e. PCA) and multivariate modeling (i.e. PLS-DA) have been used to analyse the differences between the groups of study and to select the most relevant metabolites that explain such differences. PCA, as an unsupervised method, poses little controversy towards its use and interpretation. Hence, PCA was used as a data exploration tool during the analysis and not for metabolite selection.

However, supervised methods such as PLS-DA use information about the groups to maximise differences producing a predictive model. Although used as a technique to show differences between treatments, the resulting model can also be used to assign new data samples to groups in the data that the model was built on. The latter has not been used in this thesis. However, the validity of the models needs to be considered when interpreting the results.

The number of variates that these models are built on can vary. Typically, use of more variates increases the accuracy. However, more noise associated with the data is also incorporated to the model, causing overfitting. Thus, it is imperative to use cross-validation methods to check and minimise overfitting during the model building process. The ideal approach uses rounds of validation and testing with different parts of the data: splitting of data into three parts (typically 60%, 20% and 20%): 60% used to train a model, 20% to validate it and the last 20%, never used to build any model, to test the model. Although the ideal, this approach requires a large amount of data which is unfeasible for this study. Thus, for this particular project, the data was split in a 70% and 30% manner as train and test data. Cross-validation was used to select the optimal number of variables to retain for the model using all the data. Data availability is a common shortcoming in biological studies and as such there are multiple examples in the literature in which this approach has been undertaken [238], [239].

Another limitation to discuss includes the choice of modelling approach. PLS-DA is the predominant method used to analyse metabolomics datasets due to its availability in different statistical packages, the ability to analyse highly collinear data and its ease of interpretation, relatable to original biological entities. However, it might not necessarily be the most adequate method [240]. Other common methods applied are random forest (RF) and support vector machine (SVM). Both statistical approaches are great and have their advantages. For example, RF is based on decision trees which are very intuitive to understand, and RF is a large number of decision trees created from randomly selected

variables (giving the name random forest). When RF models used in prediction each decision tree proposes a classification for the input data and the final result is generated through a voting process. This approach brings an inherited randomisation to the model making it more resistant to over-fitting of the data. On the down side, highly collinear data (such as NMR data) can introduce an unintentional bias during the randomised decision tree building process. In order to avoid such incident number of decision trees can be increased but this would come with a cost of computation, and harder interpretation of the model. Due to the nature of the model it may be more suitable for MS data where collinearity exist to a lesser degree. Ideally, new statistical methods, still to be developed, would address the typical problems in metabolomics research such as lower number of samples than variables in the dataset and allow for some of the above limitations to be addressed. In this thesis, PLS-DA was used because of all of the above advantages, coupled with relatively good results that can be contextualised within the literature.

In this project, data acquisition was performed mainly *via* ^{1}H-NMR where, depending on the molecular structure, a metabolite can be represented by a single peak or more. This property of NMR can sometimes be a limitation as well as an advantage. Multiple signals representing a molecule give confidence and redundancy in data. On the other hand, selecting important/influential features from statistical models becomes more complicated. In this project, a scoring system was created based on the correlation of NMR signals, termed the correlation reliability score (CRS). Using this scoring system, the best representative peak for the molecule in the analysis was selected. A caveat to this scoring system is that it is only useful for molecules with multiple peaks that are assigned to a molecule. This is because singlets will always score 100% correlation with themselves and it is virtually impossible to deconvolute a molecule's NMR signature solely based on correlation in a complex mixture.

6.2.6 Pathway analysis

Metabolomics studies often search for changes in pathways, typically performed by various methods collectively called pathway analyses. The majority of these pathway methods, especially of the enrichment type, use Fisher's exact test. In a pathway analysis context, this test calculates the probability of a set of selected metabolites belonging to another set of metabolites of a pathway. The resulting probabilities are then compared to a significance threshold, typically α=0.05. The results deemed significant are then accepted as significantly changing pathways. Unfortunately, this general practice can often be misleading. Any form

of pathway analysis based on Fisher's exact test or other network methods, such as betweenness centrality, can only give clues about a pathway. Furthermore, due to the highly dynamic nature of metabolic pathways, the quantification of a pathway down to a single value in order to use it in comparisons to be significantly higher or lower is not appropriate. Results obtained from a metabolite set enrichment analysis (MSEA) should be taken as leads to follow up on. If the goal of the study is to explore and discover pathways in fine detail, methods such as metabolic flux analysis or a combination of *in vitro* and *in vivo* biochemical experiments to show evidence for particular reactions would be better suited. In this project, MSEA was used as a lead as to which pathways might be affected under given experimental conditions. The final conclusions were drawn from the data presented and information available in the literature and not solely on MSEA results.

6.3 Contributions to the field

6.3.1 Bin selection for NMR data

In metabolomics research, variable selection from a statistical model is a cornerstone of the method. In mass spectrometry platforms, this can be achieved directly from the model with variable selection methods (e.g. variable importance of the projection). Due to the nature of NMR data, a metabolite can be represented by multiple bins depending on the molecular structure. When a variable selection method is applied, metabolites with multiple bins often contribute with multiple variates. In order to simplify the data, a representative bin is required. Currently, there is no method to account for this systematically. In this work, a scoring system, CRS, was developed. CRS is calculated from the correlation of a metabolite bin to the remaining bins of the same metabolite. This provides a systematic and objective approach to select a representative bin for a metabolite with multiple bin signature (i.e. multiple peaks observable in the 1D ^{1}H-NMR spectrum).

6.3.2 Biomarker candidates for sex difference in *An. gambiae* and *Ae. aegypti* targeted at energy and energy storage mechanisms

By comparing the metabolic profiles of male and female mosquitoes in pupal and adult stages, a set of metabolites were identified that were consistently distinct across different stages and species. These metabolites, namely glucose, propionate, lactate, and acetate, predominantly take part in energy and energy storage mechanisms in mosquitoes. As hypothesised, female mosquitoes require higher levels of energy in order to meet the high energy requirements for finding a blood meal, oogenesis, and oviposition. Furthermore,

these metabolites can be optimised to give an accurate identification of sex through metabolic profiling. As for application, such information can be integrated in large scale studies of a metabolomics nature where great numbers of mosquitoes can be collected and extracted rapidly, and sexing can be integrated in to the biochemical analysis. Additionally, it can be used as a secondary confirmation for the traditional sexing of mosquitoes or elucidating sample sex information for extracted legacy samples.

6.3.3 Cyp4g16 and Cyp4g17 catalyse predominantly branched alkanes and 2-methylbranched alkanes in *An. gambiae*.

Prior to this study, only Cyp4g16's decarbonylation function was shown through expression [75] and Cyp4g17's function was only observed through expression in *D. melanogaster* [186]. By employing NMR metabolomics, further evidence has been raised for Cyp4g16's and Cyp4g17's function in *An. gambiae*. In addition, the activity of these two enzymes in relation to developmental stage was shown (Cyp4g16: early pupa and early adult and Cyp4g17: early adult).

6.3.4 VK7 strain of *An. gambiae* show evidence of cuticular hydrocarbon resistance in metabolomics analysis

VK7 strain of *An. gambiae* has been identified to have cuticular resistance in previous studies [213], [214]. In this study, supporting information of cuticular resistance through pathways identified in resistant and susceptible species and their similarities with the pathways identified in knock-down strains of *An. gambiae* have been shown.

6.3.5 Trehalose as a biomarker candidate for pyrethroid resistant in wild type *An. gambiae* and *Ae. aegypti*

Further comparison of wild type *An. gambiae* and *Ae. aegypti* has shown trehalose as a biomarker which is consistently in higher abundance in resistant strains across pupal and adult stages compared to sensitive strains. This information can be utilised as a resistance biomarker for quick identification of resistant species. Furthermore, metabolic systems relating to trehalose can be exploited as new insecticide targets such as trehalose transporter (TreT) and trehalase enzymes (Tre1 and Tre2).

6.4 Future work

6.4.1 Metabolomics of CHC for complementary information

There are a several studies on the cuticular hydrocarbon composition of these knock-downs carried out *via* GC-MS, although none of these studies investigated the HC content within the mosquito. A metabolomics study targeting these HCs can further inform on the changes in the final products of Cyp4g16 and Cyp4g17.

6.4.2 Mosquito metabolite library

One of limitations in this project was the absence of a mosquito metabolomics library. In other fields, these libraries are extensive (e.g. HMDB [241] and livestock metabolome database (LMDB) [242]). Establishing a mosquito metabolome database to include information such as, metabolite identifications, identification method and associated data, location concentration ranges, associated pathways to existing databases, would greatly benefit entometabolomics. Since such a database will be built from ground up it is imperative to design this database with an API structure for easy integration with scripts and software.

6.4.3 Verification of biomarkers for sexing *An. gambiae* and *Ae. aegypti*

Using the biomarker candidates for sexing of *An. gambiae* and *Ae. aegypti*, a study should be carried out in order to validate and establish the concentration boundaries on field strain mosquitoes.

6.4.4 Characterisation of Cyp4g16 and Cyp4g17 decarbonylation.

With the growing evidence regarding Cyp4g16 and Cyp4g17 function in HC production, a biophysical characterisation of these enzymes would be beneficial in order to fully understand their roles and importance in CHC biosynthesis. A detailed pathway analysis could be further elucidated using a multi-omics approach combining transcriptomics and proteomics information. It should be noted that this multi-omics approach would be costly and considering the low concentrations, a different experimental design involving pooled cohorts could be explored.

Furthermore, elucidating the molecular structures of these enzymes would contribute a great deal of insight to the biophysics characterisation process, especially in identifying the associated potential ligands. Starting with Cyp4g16 is recommended since its expression has shown it to be relatively easier compared with Cyp4g17 [75]. Cyp4g16 and Cyp4g17 are

approximately 64 KDa so an approach utilising X-ray crystallography and/or homology modelling would be most appropriate.

6.4.5 Testing and verification of the potential resistance biomarker trehalose

This study proposed trehalose (higher in resistant strains) as a biomarker for resistance. A follow up pilot study to validate the trehalose as a biomarker needs to be conducted. This study should be designed as a field study and aim to establish the necessary trehalose measurement and threshold (trehalose per body mass or trehalose ratio to another metabolite) for identification of resistant strains. Additionally, a broader verification of trehalose as a resistance biomarker should be performed on the R and S forms of *An. gambiae*. If this study requires a secondary metabolite for normalisation, three possible metabolites can be proposed from this thesis. Changes in these metabolites were not significant and not selected by the PLS-DA models as discriminating: xanthine (adults only), succinate (pupa only) and glutamine(both).

6.5 Final remarks

In this novel study, NMR metabolomics' capabilities were pushed by investigating insecticide resistance in mosquitoes. A pipeline was built and used in the metabolic profiling of different sexes, knock-downs (Cyp4g16 & Cyp4g17), and different species of resistant and susceptible mosquitoes. This pipeline focused on analysis of polar metabolites of a relative quantitation nature led to the development of CRS. A tool to select a representative bin from multiple NMR signal, thus simplifying the analysis data. Furthermore, despite the encountered low sample number and identified metabolites evidence on the function of Cyp4g17 as a decarbonylase was shown. Additionally, sex-specific metabolic differences were reported. Lastly, for the first time resistant and susceptible mosquito species' polar metabolome was compared presenting consistent trends between resistance statuses.

7 References

[1] C. Gillott, *Entomology*, 3rd ed. Springer, 2005.

[2] N. Becker *et al.*, *Mosquitoes and Their Control*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[3] C. R. Rutledge, "Mosquitoes (Diptera: Culicidae)," in *Encyclopedia of Entomology*, Dordrecht: Springer Netherlands, 2008, pp. 2476–2483.

[4] World Health Organization and WHO, *Global vector control response 2017–2030*. World Health Organization, 2017.

[5] World Health Organization, "WHO | World malaria report 2017," *Who*, 2018.

[6] J. A. Pérez-Molina and I. Molina, "Chagas disease," *Lancet*, vol. 391, pp. 82–94, 2018.

[7] S. Burza, S. L. Croft, and M. Boelaert, "Leishmaniasis," *Lancet*, vol. 392, no. 10151, pp. 951–970, 2018.

[8] D. J. Gray, A. G. Ross, Y. S. Li, and D. P. McManus, "Diagnosis and management of schistosomiasis," *Bmj*, vol. 342, no. 7807, p. d2651, May 2011.

[9] S. Leta, T. J. Beyene, E. M. De Clercq, K. Amenu, M. U. G. Kraemer, and C. W. Revie, "Global risk mapping for major diseases transmitted by Aedes aegypti and Aedes albopictus," *Int. J. Infect. Dis.*, vol. 67, pp. 25–35, Feb. 2018.

[10] M. J. Griffiths *et al.*, "The functional, social and economic impact of acute encephalitis syndrome in Nepal--a longitudinal follow-up study.," *PLoS Negl. Trop. Dis.*, vol. 7, no. 9, p. e2383, 2013.

[11] A. D. T. Barrett, "Economic burden of West Nile virus in the United States.," *Am. J. Trop. Med. Hyg.*, vol. 90, no. 3, pp. 389–390, Mar. 2014.

[12] L. M. Barber, J. J. Schleier, R. K. D. Peterson, and R. K. D. Peterson, "Economic cost analysis of West Nile virus outbreak, Sacramento County, California, USA, 2005.," *Emerg. Infect. Dis.*, vol. 16, no. 3, pp. 480–6, Mar. 2010.

[13] A. Zohrabian *et al.*, "West Nile Virus Economic Impact, Louisiana, 2002," *Emerg. Infect. Dis.*, vol. 10, no. 10, pp. 1736–1744, Oct. 2004.

[14] J. A. Suaya *et al.*, "Cost of dengue cases in eight countries in the Americas and asia: A prospective study," *Am. J. Trop. Med. Hyg.*, vol. 80, no. 5, pp. 846–855, May 2009.

[15] D. S. Shepard, L. Coudeville, Y. A. Halasa, B. Zambrano, and G. H. Dayan, "Economic impact of dengue illness in the Americas.," *Am. J. Trop. Med. Hyg.*, vol. 84, no. 2, pp. 200–7, Feb. 2011.

[16] G. O. Muga, W. Onyango-Ouma, R. Sang, and H. Affognon, "Sociocultural and economic dimensions of Rift Valley fever.," *Am. J. Trop. Med. Hyg.*, vol. 92, no. 4, pp.

730–8, Apr. 2015.

[17] World Health Organization, “Yellow Fever Initiative - Providing an opportunity of a lifetime,” 2010.

[18] M. S. Hossain *et al.*, “Chikungunya outbreak (2017) in Bangladesh: Clinical profile, economic impact and quality of life during the acute phase of the disease.,” *PLoS Negl. Trop. Dis.*, vol. 12, no. 6, p. e0006561, 2018.

[19] C. C. Ezenduka, D. R. Falleiros, and B. B. Godman, “Evaluating the Treatment Costs for Uncomplicated Malaria at a Public Healthcare Facility in Nigeria and the Implications,” *PharmacoEconomics - Open*, vol. 1, no. 3, pp. 185–194, Sep. 2017.

[20] A. Chandy, A. S. Thakur, M. P. Singh, and A. Manigauha, “A review of neglected tropical diseases: filariasis,” *Asian Pac. J. Trop. Med.*, vol. 4, no. 7, pp. 581–586, Jul. 2011.

[21] E. A. Ottesen, “Lymphatic Filariasis: Treatment, Control and Elimination,” *Adv. Parasitol.*, vol. 61, pp. 395–441, Jan. 2006.

[22] G. Dakshinamoorthy, A. K. Samykutty, G. Munirathinam, M. V. Reddy, and R. Kalyanasundaram, “Multivalent fusion protein vaccine for lymphatic filariasis,” *Vaccine*, vol. 31, no. 12, pp. 1616–1622, Mar. 2013.

[23] J. V. J. Silva *et al.*, “A scoping review of Chikungunya virus infection: epidemiology, clinical characteristics, viral co-circulation complications, and control,” *Acta Trop.*, vol. 188, pp. 213–224, Dec. 2018.

[24] K. Tharmarajah, S. Mahalingam, and A. Zaid, “Chikungunya: vaccines and therapeutics.,” *F1000Research*, vol. 6, p. 2114, 2017.

[25] R. Tschismarov, A. Pfeiffer, M. Muellner, E. Tauber, K. Ramsauer, and E. C. Reisinger, “Immunogenicity, safety, and tolerability of the measles-vectored chikungunya virus vaccine MV-CHIK: a double-blind, randomised, placebo-controlled and active-controlled phase 2 trial,” *www.thelancet.com*, vol. 392, 2018.

[26] M. G. Guzman, D. J. Gubler, A. Izquierdo, E. Martinez, and S. B. Halstead, “Dengue infection,” *Nat. Rev. Dis. Prim.*, vol. 2, no. 1, p. 16055, Dec. 2016.

[27] O. J. Brady *et al.*, “Refining the Global Spatial Limits of Dengue Virus Transmission by Evidence-Based Consensus,” *PLoS Negl. Trop. Dis.*, vol. 6, no. 8, p. e1760, Aug. 2012.

[28] B. Guy, B. Barrere, C. Malinowski, M. Saville, R. Teyssou, and J. Lang, “From research to phase III: Preclinical, industrial and clinical development of the Sanofi Pasteur tetravalent dengue vaccine,” *Vaccine*, vol. 29, no. 42, pp. 7229–7241, Sep. 2011.

[29] P. Pitisuttithum and A. Bouckenooghe, “The first licensed dengue vaccine: an important tool for integrated preventive strategies against dengue virus infection,”

Expert Rev. Vaccines, vol. 15, no. 7, pp. 795–798, Jul. 2016.

[30] S. Swaminathan and N. Khanna, "Dengue vaccine development: Global and Indian scenarios," *Int. J. Infect. Dis.*, Jan. 2019.

[31] S. Lumley, D. L. Horton, L. L. M. Hernandez-Triana, N. Johnson, A. R. Fooks, and R. Hewson, "Rift valley fever virus: Strategies for maintenance, survival and vertical transmission in mosquitoes," *J. Gen. Virol.*, vol. 98, no. 5, pp. 875–887, 2017.

[32] B. Faburay, A. D. LaBeaud, D. S. McVey, W. C. Wilson, and J. A. Richt, "Current Status of Rift Valley Fever Vaccine Development.," *Vaccines*, vol. 5, no. 3, Sep. 2017.

[33] T. P. Monath and P. F. C. Vasconcelos, "Yellow fever.," *J. Clin. Virol.*, vol. 64, pp. 160–73, Mar. 2015.

[34] D. Baud, D. J. Gubler, B. Schaub, M. C. Lanteri, and D. Musso, "An update on Zika virus infection," *Lancet*, vol. 390, no. 10107, pp. 2099–2109, Nov. 2017.

[35] V. C. Agumadu and K. Ramphul, "Zika Virus: A Review of Literature.," *Cureus*, vol. 10, no. 7, p. e3025, Jul. 2018.

[36] D. Musso and D. J. Gubler, "Zika Virus.," *Clin. Microbiol. Rev.*, vol. 29, no. 3, pp. 487–524, 2016.

[37] Y. Ophir and K. H. Jamieson, "The Effects of Zika Virus Risk Coverage on Familiarity, Knowledge and Behavior in the U.S. – A Time Series Analysis Combining Content Analysis and a Nationally Representative Survey," *Health Commun.*, pp. 1–11, Oct. 2018.

[38] P. Abbink, K. E. Stephenson, and D. H. Barouch, "Zika virus vaccines," *Nat. Rev. Microbiol.*, vol. 16, no. 10, pp. 594–600, Oct. 2018.

[39] B. M. Greenwood *et al.*, "Malaria: progress, perils, and prospects for eradication," *J. Clin. Invest.*, vol. 118, no. 4, pp. 1266–1276, Apr. 2008.

[40] R. Varo *et al.*, "Adjunctive therapy for severe malaria: a review and critical appraisal," *Malar. J.*, vol. 17, no. 1, p. 47, Dec. 2018.

[41] E. A. Ashley, A. Pyae Phyo, and C. J. Woodrow, "Malaria," *Lancet*, vol. 391, no. 10130, pp. 1608–1621, 2018.

[42] A. F. Cowman, J. Healer, D. Marapana, and K. Marsh, "Leading Edge Review Malaria: Biology and Disease," *Cell*, vol. 167, pp. 610–624, 2016.

[43] B. Blasco, D. Leroy, and D. A. Fidock, "Antimalarial drug resistance: linking Plasmodium falciparum parasite biology to the clinic.," *Nat. Med.*, vol. 23, no. 8, pp. 917–928, Aug. 2017.

[44] M. A. Penny *et al.*, "Public health impact and cost-effectiveness of the RTS,S/AS01 malaria vaccine: a systematic comparison of predictions from four mathematical

models.," *Lancet (London, England)*, vol. 387, no. 10016, pp. 367–375, Jan. 2016.

[45] C. A. Dimala, B. T. Kika, B. M. Kadia, and H. Blencowe, "Current challenges and proposed solutions to the effective implementation of the RTS, S/AS01 Malaria Vaccine Program in sub-Saharan Africa: A systematic review," *PLoS One*, vol. 13, no. 12, p. e0209744, Dec. 2018.

[46] S. Mahmoudi and H. Keshavarz, "Efficacy of Phase 3 Trial of RTS, S/AS01 Malaria Vaccine in infants: a systematic review and meta-analysis," *Hum. Vaccin. Immunother.*, pp. 00–00, Jan. 2017.

[47] D. F. A. Diniz, C. M. R. de Albuquerque, L. O. Oliva, M. A. V. de Melo-Santos, and C. F. J. Ayres, "Diapause and quiescence: dormancy mechanisms that contribute to the geographical expansion of mosquitoes and their evolutionary success," *Parasit. Vectors*, vol. 10, 2017.

[48] J. F. Day, "Mosquito Oviposition Behavior and Vector Control.," *Insects*, vol. 7, no. 4, Nov. 2016.

[49] L. C. Farnesi, H. C. M. Vargas, D. Valle, and G. L. Rezende, "Darker eggs of mosquitoes resist more to dry conditions: Melanin enhances serosal cuticle contribution in egg resistance to desiccation in Aedes, Anopheles and Culex vectors," *PLoS Negl. Trop. Dis.*, vol. 11, no. 10, p. e0006063, Oct. 2017.

[50] J. E. Casida, "Pesticide Interactions: Mechanisms, Benefits, and Risks," *J. Agric. Food Chem.*, vol. 65, no. 23, pp. 4553–4561, Jun. 2017.

[51] T. G. E. Davies, L. M. Field, P. N. R. Usherwood, and M. S. Williamson, "DDT, pyrethrins, pyrethroids and insect sodium channels," *IUBMB Life*, vol. 59, no. 3, pp. 151–162, 2007.

[52] C. Antonio-Nkondjio *et al.*, "Review of the evolution of insecticide resistance in main malaria vectors in Cameroon from 1990 to 2017," *Parasit. Vectors*, vol. 10, no. 1, p. 472, Dec. 2017.

[53] J. L. Capinera *et al.*, "Neurological Effects of Insecticides and the Insect Nervous System," in *Encyclopedia of Entomology*, Dordrecht: Springer Netherlands, 2008, pp. 2596–2605.

[54] T. R. Fukuto, "Mechanism of action of organophosphorus and carbamate insecticides.," *Environ. Health Perspect.*, vol. 87, pp. 245–254, Jul. 1990.

[55] N. Simon-Delso *et al.*, "Systemic insecticides (neonicotinoids and fipronil): trends, uses, mode of action and metabolites," *Environ. Sci. Pollut. Res.*, vol. 22, no. 1, pp. 5–34, Jan. 2015.

[56] B. Tanna, W. Welch, L. Ruest, J. L. Sutko, and A. J. Williams, "The interaction of a

neutral ryanoid with the ryanodine receptor channel provides insights into the mechanisms by which ryanoid binding is modulated by voltage.," *J. Gen. Physiol.*, vol. 116, no. 1, pp. 1–9, Jul. 2000.

[57] M. Gauthier, "State of the art on insect nicotinic acetylcholine receptor function in learning and memory.," *Adv. Exp. Med. Biol.*, vol. 683, pp. 97–115, 2010.

[58] S. C. R. Lummis and D. B. Sattelle, "Insect central nervous system γ-aminobutyric acid," *Neurosci. Lett.*, vol. 60, no. 1, pp. 13–18, Sep. 1985.

[59] M. Porta, P. L. Diaz-Sylvester, A. Nani, J. Ramos-Franco, and J. A. Copello, "Ryanoids and imperatoxin affect the modulation of cardiac ryanodine receptors by dihydropyridine receptor Peptide A," *Biochim. Biophys. Acta - Biomembr.*, vol. 1778, no. 11, pp. 2469–2479, Nov. 2008.

[60] P. N. R. Usherwood and H. Vais, "Towards the development of ryanoid insecticides with low mammalian toxicity," *Toxicol. Lett.*, vol. 82–83, pp. 247–254, Dec. 1995.

[61] World Health Organization, "Test procedures for insecticide resistance monitoring in malaria vector mosquitoes Global Malaria Programme," 2018.

[62] W. S. Abbott, "A Method of Computing the Effectiveness of an Insecticide," *J. Econ. Entomol.*, vol. 18, no. 2, pp. 265–267, Apr. 1925.

[63] J. A. Gómez-Guzmán, F. J. García-Marín, M. Sáinz-Pérez, and R. González-Ruiz, "Behavioural Resistance in Insects: Its Potential Use as Bio Indicator of Organic Agriculture," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 95, no. 4, p. 042038, Dec. 2017.

[64] C. Sokhna, M. O. Ndiath, and C. Rogier, "The changes in mosquito vector behaviour and the emerging resistance to insecticides will challenge the decline of malaria," *Clin. Microbiol. Infect.*, vol. 19, no. 10, pp. 902–907, Oct. 2013.

[65] E. K. Thomsen *et al.*, "Mosquito Behavior Change After Distribution of Bednets Results in Decreased Protection Against Malaria Exposure.," *J. Infect. Dis.*, vol. 215, no. 5, pp. 790–797, 2017.

[66] M. L. Gatton *et al.*, "The importance of mosquito behavioural adaptations to malaria control in Africa.," *Evolution*, vol. 67, no. 4, pp. 1218–30, Apr. 2013.

[67] C. L. Moyes *et al.*, "Contemporary status of insecticide resistance in the major Aedes vectors of arboviruses infecting humans," *PLoS Negl. Trop. Dis.*, vol. 11, no. 7, p. e0005625, Jul. 2017.

[68] R. H. Ffrench-Constant, B. Pittendrigh, A. Vaughan, and N. Anthony, "Why are there so few resistance-associated mutations in insecticide target genes?," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 353, no. 1376, pp. 1685–1693, 1998.

[69] M. Auteri, F. La Russa, V. Blanda, and A. Torina, " Insecticide Resistance Associated

with kdr Mutations in Aedes albopictus : An Update on Worldwide Evidences ," *Biomed Res. Int.*, vol. 2018, pp. 1–10, Aug. 2018.

[70] R. H. Ffrench-Constant, "The molecular genetics of insecticide resistance," *Genetics*, vol. 194, no. 4, pp. 807–815, Aug. 2013.

[71] N. Liu, "Insecticide Resistance in Mosquitoes: Impact, Mechanisms, and Research Directions," *Annu. Rev. Entomol.*, vol. 60, no. 1, pp. 537–559, Jan. 2015.

[72] N. H. Z. Safi *et al.*, "Evidence of metabolic mechanisms playing a role in multiple insecticides resistance in Anopheles stephensi populations from Afghanistan.," *Malar. J.*, vol. 16, no. 1, p. 100, 2017.

[73] J. B. Heppner *et al.*, "Integument: Structure and Function," in *Encyclopedia of Entomology*, Dordrecht: Springer Netherlands, 2008, pp. 2015–2019.

[74] V. Balabanidou, L. Grigoraki, and J. Vontas, "Insect cuticle: a critical determinant of insecticide resistance This review comes from a themed issue on Pests and resistance," *Curr. Opin. Insect Sci.*, vol. 27, pp. 68–74, Jun. 2018.

[75] V. Balabanidou *et al.*, " Cytochrome P450 associated with insecticide resistance catalyzes cuticular hydrocarbon production in Anopheles gambiae ," *Proc. Natl. Acad. Sci.*, vol. 113, no. 33, pp. 9268–9273, Aug. 2016.

[76] G. A. Yahouédo *et al.*, "Contributions of cuticle permeability and enzyme detoxification to pyrethroid resistance in the major malaria vector Anopheles gambiae.," *Sci. Rep.*, vol. 7, no. 1, p. 11091, 2017.

[77] V. A. Ingham, "Identification of novel transcripts involved in insecticide resistance in African malaria vectors," Liverpool School of Tropical Medicine, 2015.

[78] G. J. Blomquist and A.-G. Bagnères, *Insect hydrocarbons : biology, biochemistry, and chemical ecology*. Cambridge University Press, 2010.

[79] J.-M. Jallon and C. Wicker-Thomas, "Genetic studies on pheromone production in Drosophila," *Insect Pheromone Biochem. Mol. Biol.*, pp. 253–281, Jan. 2003.

[80] F. C. Ingleby, "Insect Cuticular Hydrocarbons as Dynamic Traits in Sexual Communication.," *Insects*, vol. 6, no. 3, pp. 732–42, Aug. 2015.

[81] S. D. Leonhardt, F. Menzel, V. Nehring, and T. Schmitt, "Leading Edge Review Ecology and Evolution of Communication in Social Insects," *Cell*, vol. 164, pp. 1277–1287, 2016.

[82] M. I. Stefana *et al.*, "Developmental diet regulates Drosophila lifespan via lipid autotoxins."

[83] A. M. Al Ahmed, A.-Y. Badjah-Hadj-Ahmed, Z. A. Al Othman, and M. F. Sallam, "Identification of wild collected mosquito vectors of diseases using gas

chromatography-mass spectrometry in Jazan Province, Saudi Arabia," *J. Mass Spectrom.*, vol. 48, no. 11, pp. 1170–1177, Nov. 2013.

[84] G. I. Anyanwu, D. H. Molyneux, and A. Phillips, "Variation in Cuticular Hydrocarbons among Strains of the Anopheles gambiae sensu stricto by Analysis of Cuticular Hydrocarbons Using Gas Liquid Chromatography of Larvae," *Mem. Inst. Oswaldo Cruz*, vol. 95, no. 3, pp. 295–300, 2000.

[85] A. R. Polerstock, S. D. Eigenbrode, and M. J. Klowden, "Mating Alters the Cuticular Hydrocarbons of Female *Anopheles gambiae* sensu stricto and *Aedes aegypti* (Diptera: Culicidae)," *J. Med. Entomol.*, vol. 39, no. 3, pp. 545–552, May 2002.

[86] H. Chung and S. B. Carroll, "Wax, sex and the origin of species: Dual roles of insect cuticular hydrocarbons in adaptation and mating," *BioEssays*, vol. 37, no. 7, pp. 822–830, Jul. 2015.

[87] Y. Qiu *et al.*, "An insect-specific P450 oxidative decarbonylase for cuticular hydrocarbon biosynthesis.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 37, pp. 14858–63, Sep. 2012.

[88] R. W. Howard and G. J. Blomquist, "ECOLOGICAL, BEHAVIORAL, AND BIOCHEMICAL ASPECTS OF INSECT HYDROCARBONS," *Annu. Rev. Entomol.*, vol. 50, no. 1, pp. 371–393, Jan. 2005.

[89] T. P. Chan Yong and J.-M. Jallon, "Synthèse de novo d'hydrocarbures potentiellement aphrodisiaques chez les Drosophiles," *Comptes rendus l'Académie des Sci. Paris, Série III*, vol. 303, pp. 197–202, 1986.

[90] F. P. Drijfhout, R. Kather, and S. J. Martin, "The Role of Cuticular Hydrocarbons in Insects," *Behav. Chem. Ecol.*, pp. 91–114, 2010.

[91] H. G. N. Khoo and Kim Ping Wong, "Acetyl CoA generation and N-acetylation of serotonin (5HT) in the mosquito, Aedes togoi," *Insect Biochem. Mol. Biol.*, vol. 24, no. 5, pp. 445–451, 1994.

[92] A. Alabaster *et al.*, "Deficiencies in acetyl-CoA carboxylase and fatty acid synthase 1 differentially affect eggshell formation and blood meal digestion in Aedes aegypti," *Insect Biochem. Mol. Biol.*, vol. 41, no. 12, pp. 946–955, Dec. 2011.

[93] P. W. Wertz, *Waxes: chemistry, molecular biology and functions*, vol. 79, no. 2. 2003.

[94] G. J. Blomquist and L. L. Jackson, "Chemistry and biochemistry of insect waxes," *Prog. Lipid Res.*, vol. 17, no. 4, pp. 319–345, Jan. 1979.

[95] P. R. O. de Montellano, *Cytochrome P450: Structure, mechanism, and biochemistry*, 4th ed. 2015.

[96] R. Makki, E. Cinnamon, and A. P. Gould, "The Development and Functions of

Oenocytes," *Annu. Rev. Entomol.*, vol. 59, no. 1, pp. 405–425, Jan. 2014.

[97] E. G. Hrycay and S. M. Bandiera, *The monooxygenase, peroxidase, and peroxygenase properties of cytochrome P450*, vol. 522, no. 2. 2012.

[98] G. A. Roberts, G. Grogan, A. Greter, S. L. Flitsch, and N. J. Turner, "Identification of a new class of cytochrome P450 from a Rhodococcus sp.," *J. Bacteriol.*, vol. 184, no. 14, pp. 3898–908, Jul. 2002.

[99] N. Liu, T. Li, W. R. Reid, T. Yang, and L. Zhang, "Multiple Cytochrome P450 Genes: Their Constitutive Overexpression and Permethrin Induction in Insecticide Resistant Mosquitoes, Culex quinquefasciatus," *PLoS One*, vol. 6, no. 8, p. e23403, Aug. 2011.

[100] M. A. Riaz, R. Poupardin, S. Reynaud, C. Strode, H. Ranson, and J. P. David, "Impact of glyphosate and benzo[a]pyrene on the tolerance of mosquito larvae to chemical insecticides. Role of detoxification genes in response to xenobiotics," *Aquat. Toxicol.*, vol. 93, no. 1, pp. 61–69, Jun. 2009.

[101] R. Poupardin, M. A. Riaz, J. Vontas, J. P. David, and S. Reynaud, "Transcription profiling of eleven cytochrome P450s potentially involved in xenobiotic metabolism in the mosquito *Aedes aegypti*," *Insect Mol. Biol.*, vol. 19, no. 2, pp. 185–193, Apr. 2010.

[102] R. Poupardin, S. Reynaud, C. Strode, H. Ranson, J. Vontas, and J.-P. David, "Cross-induction of detoxification genes by environmental xenobiotics and insecticides in the mosquito Aedes aegypti: Impact on larval tolerance to chemical insecticides," *Insect Biochem. Mol. Biol.*, vol. 38, no. 5, pp. 540–551, May 2008.

[103] J. Vontas *et al.*, "Gene expression in insecticide resistant and susceptible Anopheles gambiae strains constitutively or after insecticide exposure," *Insect Mol. Biol.*, vol. 14, no. 5, pp. 509–521, 2005.

[104] M. A. Schuler and M. R. Berenbaum, "Structure and Function of Cytochrome P450S in Insect Adaptation to Natural and Synthetic Toxins: Insights Gained from Molecular Modeling," *J. Chem. Ecol.*, vol. 39, no. 9, pp. 1232–1245, Sep. 2013.

[105] J.-P. David, H. M. Ismail, A. Chandor-Proust, and M. J. I. Paine, "Role of cytochrome P450s in insecticide resistance: impact on the control of mosquito-borne diseases and use of insecticides on Earth," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 368, no. 1612, pp. 1–12, Jan. 2013.

[106] R. Feyereisen, "Insect CYP Genes and P450 Enzymes," in *Insect Molecular Biology and Biochemistry*, Elsevier, 2012, pp. 236–316.

[107] T. Yang and N. Liu, "Genome analysis of cytochrome P450s and their expression profiles in insecticide resistant mosquitoes, Culex quinquefasciatus.," *PLoS One*, vol. 6, no. 12, p. e29418, 2011.

[108] I. H. Ishak *et al.*, "The Cytochrome P450 gene CYP6P12 confers pyrethroid resistance in kdr-free Malaysian populations of the dengue vector Aedes albopictus," *Sci. Rep.*, vol. 6, no. 1, p. 24707, Jul. 2016.

[109] P. Müller *et al.*, "Field-Caught Permethrin-Resistant Anopheles gambiae Overexpress CYP6P3, a P450 That Metabolises Pyrethroids," *PLoS Genet.*, vol. 4, no. 11, p. e1000286, Nov. 2008.

[110] L. Gilbert, *Insect Molecular Biology and Biochemistry*. Elsevier, 2012.

[111] H. Ranson *et al.*, "Molecular analysis of multiple cytochrome P450 genes from the malaria vector, Anopheles gambiae," *Insect Mol. Biol.*, vol. 11, no. 5, pp. 409–418, 2002.

[112] P. Pignatelli, V. A. Ingham, V. Balabanidou, J. Vontas, G. Lycett, and H. Ranson, "The *Anopheles gambiae* ATP-binding cassette transporter family: phylogenetic analysis and tissue localization provide clues on function and role in insecticide resistance," *Insect Mol. Biol.*, vol. 27, no. 1, pp. 110–122, Feb. 2018.

[113] F. Savarit, G. Sureau, M. Cobb, and J. F. Ferveur, "Genetic elimination of known pheromones reveals the fundamental chemical bases of mating and isolation in Drosophila.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 16, pp. 9015–20, Aug. 1999.

[114] H. Chung, T. Sztal, S. Pasricha, M. Sridhar, P. Batterham, and P. J. Daborn, "Characterization of Drosophila melanogaster cytochrome P450 genes," *Proc. Natl. Acad. Sci.*, vol. 106, no. 14, pp. 5731–5736, Apr. 2009.

[115] E. Gutierrez, D. Wiggins, B. Fielding, and A. P. Gould, "Specialized hepatocyte-like cells regulate Drosophila lipid metabolism," *Nature*, vol. 445, no. 7125, pp. 275–280, Jan. 2007.

[116] M. R. Viant, I. J. Kurland, M. R. Jones, and W. B. Dunn, "How close are we to complete annotation of metabolomes?," *Curr. Opin. Chem. Biol.*, vol. 36, pp. 64–69, Feb. 2017.

[117] C. B. Newgard, "Metabolomics and Metabolic Diseases: Where Do We Stand?," *Cell Metab.*, vol. 25, no. 1, pp. 43–56, 2017.

[118] S. Hayton, G. L. Maker, I. Mullaney, and R. D. Trengove, "Experimental design and reporting standards for metabolomics studies of mammalian cell lines," *Cell. Mol. Life Sci.*, vol. 74, no. 24, pp. 4421–4441, Dec. 2017.

[119] D. S. Wishart *et al.*, "HMDB 4.0: The human metabolome database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D608–D617, Jan. 2018.

[120] K. Bingol, "Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods.," *High-throughput*, vol. 7, no. 2, Apr. 2018.

[121] A. Marco-Ramell *et al.*, "Evaluation and comparison of bioinformatic tools for the

enrichment analysis of metabolomics data," *BMC Bioinformatics*, vol. 19, no. 1, p. 1, Dec. 2018.

[122] G. A. N. Gowda and D. Djukovic, "Overview of Mass Spectrometry-Based Metabolomics: Opportunities and Challenges," *Methods Mol Biol*, vol. 1198, pp. 3–12, 2014.

[123] J. Keeler, "Understanding NMR Spectroscopy," 2010.

[124] T. D. Lash^ and S. S. Lash, "The Use of Pascal-like Triangles in Describing First-Order NMR Coupling Patterns Similarly the triangle for boron-11 couplings3 (7 = %) (see reference 7) may be generated from the symmetric expres-sion (x3 + x2y + xy2 + y3)rt. Multiplets due to coupling with nuclei with higher spin values may be described by the fol-lowing expressions (numbers at right refer to references be-low). Relative intensities 7 = 2 = % (x4 + x3y + x2y2 + xy3 + y4)n."

[125] J. Aires-de-Sousa, M. C. Hemmer, and J. Gasteiger, "Prediction of 1H NMR chemical shifts using neural networks," *Anal. Chem.*, vol. 74, no. 1, pp. 80–90, Jan. 2002.

[126] A. M. Castillo, L. Patiny, and J. Wist, "Fast and accurate algorithm for the simulation of NMR spectra of large spin systems," *J. Magn. Reson.*, vol. 209, no. 2, pp. 123–130, Apr. 2011.

[127] D. Banfi and L. Patiny, "<I>www.nmrdb.org</I>: Resurrecting and Processing NMR Spectra On-line," *Chim. Int. J. Chem.*, vol. 62, no. 4, pp. 280–281, Apr. 2008.

[128] P. Soininen *et al.*, "High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism," *Analyst*, vol. 134, no. 9, pp. 1781–1785, Sep. 2009.

[129] A.-H. Emwas *et al.*, "NMR Spectroscopy for Metabolomics Research," *Metabolites*, vol. 9, no. 7, p. 123, Jun. 2019.

[130] C. K. Larive, G. A. Barding, and M. M. Dinges, "NMR Spectroscopy for Metabolomics and Metabolic Profiling," *Anal. Chem.*, vol. 87, no. 1, pp. 133–146, Jan. 2015.

[131] S. Y. Lee, J. M. Park, and T. Y. Kim, "Application of Metabolic Flux Analysis in Metabolic Engineering," *Methods Enzymol.*, vol. 498, pp. 67–93, Jan. 2011.

[132] G. Theodoridis, H. Gika, and W. Ian, *Metabolic Profiling*, vol. 1738. New York, NY: Springer New York, 2018.

[133] S. Zhang, G. A. Nagana Gowda, T. Ye, and D. Raftery, "Advances in NMR-based biofluid analysis and metabolite profiling," *Analyst*, vol. 135, no. 7, pp. 1490–1498, 2010.

[134] M.-E. P. Papadimitropoulos, C. G. Vasilopoulou, C. Maga-Nteve, and M. I. Klapa, "Untargeted GC-MS Metabolomics," in *Methods in molecular biology (Clifton, N.J.)*, vol. 1738, 2018, pp. 133–147.

[135] A. Scalbert *et al.*, “Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research.,” *Metabolomics*, vol. 5, no. 4, pp. 435–458, Dec. 2009.

[136] C. Pan, X. Xu, H. He, X. Cai, and X. Zhang, “Separation and identification of cis and trans isomers of 2-butene-1,4-diol and lafutidine by HPLC and LC-MS.,” *J. Zhejiang Univ. Sci. B*, vol. 6, no. 1, pp. 74–8, Jan. 2005.

[137] M. Belka, W. Hewelt-Belka, J. Sławiński, and T. Bączek, “Mass spectrometry based identification of geometric isomers during metabolic stability study of a new cytotoxic sulfonamide derivatives supported by quantitative structure-retention relationships.,” *PLoS One*, vol. 9, no. 6, p. e98096, 2014.

[138] D. Broadhurst *et al.*, “Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies,” *Metabolomics*, vol. 14, no. 6, p. 72, 2018.

[139] E. Zelena *et al.*, “Development of a robust and repeatable UPLC - MS method for the long-term metabolomic study of human serum,” *Anal. Chem.*, vol. 81, no. 4, pp. 1357–1364, Feb. 2009.

[140] S. Wernisch and S. Pennathur, “Evaluation of coverage, retention patterns, and selectivity of seven liquid chromatographic methods for metabolomics,” *Anal. Bioanal. Chem.*, vol. 408, no. 22, pp. 6079–6091, Sep. 2016.

[141] E. L. Ulrich *et al.*, “BioMagResBank,” *Nucleic Acids Res.*, vol. 36, pp. D402-8, Jan. 2008.

[142] M. j. Nueda, A. Ferrer, and A. Conesa, “ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments,” *Biostatistics*, vol. 13, no. 3, pp. 553–566, Jul. 2012.

[143] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods,” *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007.

[144] J. T. Leek and J. D. Storey, “Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis,” *PLoS Genet.*, vol. 3, no. 9, p. e161, 2007.

[145] J. Li, P. R. Bushel, T. M. Chu, and R. D. Wolfinger, “Principal Variance Components Analysis: Estimating Batch Effects in Microarray Gene Expression Data,” in *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*, Chichester, UK: John Wiley & Sons, Ltd, 2009, pp. 141–154.

[146] J. Chong *et al.*, “MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis,” *Nucleic Acids Res.*, vol. 46, no. W1, pp. W486–W494, Jul. 2018.

[147] M. I. Mhlongo, L. A. Piater, N. E. Madala, N. Labuschagne, and I. A. Dubery, “The Chemistry of Plant–Microbe Interactions in the Rhizosphere and the Potential for Metabolomics to Reveal Signaling Related to Defense Priming and Induced Systemic Resistance,” *Front. Plant Sci.*, vol. 9, p. 112, Feb. 2018.

[148] L. Kešnerová, R. A. T. Mars, K. M. Ellegaard, M. Troilo, U. Sauer, and P. Engel, “Disentangling metabolic functions of bacteria in the honey bee gut,” *PLOS Biol.*, vol. 15, no. 12, p. e2003467, Dec. 2017.

[149] J. C. Hoxmeier *et al.*, “Analysis of the metabolome of Anopheles gambiae mosquito after exposure to Mycobacterium ulcerans,” *Sci. Rep.*, vol. 5, no. 1, p. 9242, Mar. 2015.

[150] C. Sanchez-Arcos, M. Kai, A. Svatoš, J. Gershenzon, and G. Kunert, “Untargeted Metabolomics Approach Reveals Differences in Host Plant Chemistry Before and After Infestation With Different Pea Aphid Host Races,” *Front. Plant Sci.*, vol. 10, p. 188, Feb. 2019.

[151] D. Maag, M. Erb, and G. Glauser, “Metabolomics in plant-herbivore interactions: challenges and applications,” *Entomol. Exp. Appl.*, vol. 157, no. 1, pp. 18–29, Oct. 2015.

[152] P. Lehmann *et al.*, “Metabolome dynamics of diapause in the butterfly Pieris napi: distinguishing maintenance, termination and post-diapause phases,” 2017.

[153] K. A. Aliferis and M. Chrysayi-Tokousbalides, “Metabolomics in pesticide research and development: review and future perspectives,” *Metabolomics*, vol. 7, no. 1, pp. 35–53, Mar. 2011.

[154] K. Derecka *et al.*, “Transient Exposure to Low Levels of Insecticide Affects Metabolic Networks of Honeybee Larvae,” *PLoS One*, vol. 8, no. 7, p. e68191, Jul. 2013.

[155] A. Malmendal *et al.*, “Metabolomic profiling of heat stress: hardening and recovery of homeostasis in *Drosophila*,” *Am. J. Physiol. Integr. Comp. Physiol.*, vol. 291, no. 1, pp. R205–R212, Jul. 2006.

[156] T. C. Hawes, A. C. Hines, M. R. Viant, J. S. Bale, M. R. Worland, and P. Convey, “Metabolomic fingerprint of cryo-stress in a freeze tolerant insect.,” *Cryo Letters*, vol. 29, no. 6, pp. 505–15, 2008.

[157] C. J. P. Snart, I. C. W. Hardy, and D. A. Barrett, “Entometabolomics: applications of modern analytical techniques to insect studies.,” *Entomol. Exp. Appl.*, vol. 155, no. 1, pp. 1–17, Apr. 2015.

[158] J. E. Cox, C. S. Thummel, and J. M. Tennessen, “Metabolomic Studies in Drosophila.,” *Genetics*, vol. 206, no. 3, pp. 1169–1185, Jul. 2017.

[159] D. P. Price, F. D. Schilkey, A. Ulanov, and I. A. Hansen, “Small mosquitoes, large

implications: crowding and starvation affects gene expression and nutrient accumulation in Aedes aegypti," *Parasit. Vectors*, vol. 8, no. 1, p. 252, Dec. 2015.

[160] J. Shrinet, N. S. Bhavesh, and S. Sunil, "Understanding Oxidative Stress in Aedes during Chikungunya and Dengue Virus Infections Using Integromics Analysis.," *Viruses*, vol. 10, no. 6, 2018.

[161] Y. Hou *et al.*, "Temporal Coordination of Carbohydrate Metabolism during Mosquito Reproduction," *PLOS Genet.*, vol. 11, no. 7, p. e1005309, Jul. 2015.

[162] S. M. Prud'homme, D. Renault, J.-P. David, and S. Reynaud, "Multiscale Approach to Deciphering the Molecular Mechanisms Involved in the Direct and Intergenerational Effect of Ibuprofen on Mosquito *Aedes aegypti*," *Environ. Sci. Technol.*, vol. 52, no. 14, pp. 7937–7950, Jul. 2018.

[163] C. Rivera-Perez, M. Nouzova, I. Lamboglia, and F. G. Noriega, "Metabolic analysis reveals changes in the mevalonate and juvenile hormone synthesis pathways linked to the mosquito reproductive physiology," *Insect Biochem. Mol. Biol.*, vol. 51, pp. 1–9, Aug. 2014.

[164] T. D. Horvath, S. Dagan, P. L. Lorenzi, D. H. Hawke, and P. Y. Scaraffia, "Positional stable isotope tracer analysis reveals carbon routes during ammonia metabolism of *Aedes aegypti* mosquitoes," *FASEB J.*, vol. 32, no. 1, pp. 466–477, Jan. 2018.

[165] Z. A. Batz and P. A. Armbruster, "Diapause-associated changes in the lipid and metabolite profiles of the Asian tiger mosquito, Aedes albopictus.," *J. Exp. Biol.*, vol. 221, no. Pt 24, p. jeb.189480, Dec. 2018.

[166] K. Hidalgo *et al.*, "Comparative physiological plasticity to desiccation in distinct populations of the malarial mosquito Anopheles coluzzii," *Parasit. Vectors*, vol. 9, no. 1, p. 565, Dec. 2016.

[167] K. H. Lockey, "The Thickness of Some Insect Epicuticular Wax Layers," *J. Exp. Biol.*, vol. 37, pp. 316–239, 1960.

[168] K. H. Lockey, "Insect cuticular hydrocarbons," *Comp. Biochem. Physiol. -- Part B Biochem.*, vol. 65, no. 3, pp. 457–462, 1980.

[169] K. H. Lockey, "Minireview insect cuticular lipids," *Comp. Biochem. Physiol.*, 1985.

[170] K. H. Lockey, "Cuticular hydrocarbons of locusta, schistocerca, and Periplaneta, and their role in waterproofing," *Insect Biochem.*, vol. 6, no. 5, pp. 457–472, 1976.

[171] K. H. Lockey, "Insect hydrocarbon classes: Implications for chemotaxonomy," *Insect Biochem.*, vol. 21, no. 1, pp. 91–97, 1991.

[172] K. H. Lockey, "Lipids of the insect cuticle: origin, composition and function," *Comp. Biochem. Physiol. -- Part B Biochem.*, vol. 89, no. 4, pp. 595–645, 1988.

[173] R. W. Howard, C. A. McDaniel, and G. J. Blomquist, “Chemical mimicry as an integrating mechanism: Cuticular hydrocarbons of a termitophile and its host,” *Science (80-.).*, vol. 210, no. 4468, pp. 431–433, Oct. 1980.

[174] G. J. Blomquist, D. R. Nelson, and M. De Renobales, “Chemistry, biochemistry, and physiology of insect cuticular lipids,” *Arch. Insect Biochem. Physiol.*, vol. 6, no. 4, pp. 227–265, 1987.

[175] D. W. Stanley-Samuelson, R. A. Jurenka, C. Cripps, G. J. Blomquist, and M. de Renobales, “Fatty acids in insects: Composition, metabolism, and biological significance,” *Arch. Insect Biochem. Physiol.*, vol. 9, no. 1, pp. 1–33, 1988.

[176] N. A. Hamid *et al.*, “Behavioral response of Aedes Aegypti (L.) to its semiochemicals,” *Southeast Asian J. Trop. Med. Public Health*, vol. 47, no. 4, pp. 691–700, 2016.

[177] K. Liebman, I. Swamidoss, L. Vizcaino, A. Lenhart, F. Dowell, and R. Wirtz, “The influence of diet on the use of near-infrared spectroscopy to determine the age of female aedes aegypti mosquitoes,” *Am. J. Trop. Med. Hyg.*, vol. 92, no. 5, pp. 1070–1075, May 2015.

[178] J. M. Urbanski, J. B. Benoit, M. R. Michaud, D. L. Denlinger, and P. Armbruster, “The molecular physiology of increased egg desiccation resistance during diapause in the invasive mosquito, Aedes albopictus,” *Proc. R. Soc. B Biol. Sci.*, vol. 277, no. 1694, p. 2683, Sep. 2010.

[179] A. C. Arcaz *et al.*, “Desiccation tolerance in Anopheles coluzzii: the effects of spiracle size and cuticular hydrocarbons.,” *J. Exp. Biol.*, vol. 219, no. Pt 11, pp. 1675–88, Jun. 2016.

[180] K. M. Wagoner, T. Lehmann, D. L. Huestis, B. M. Ehrmann, N. B. Cech, and G. Wasserberg, “Identification of morphological and chemical markers of dry- and wet-season conditions in female Anopheles gambiae mosquitoes,” *Parasit. Vectors*, vol. 7, no. 1, p. 294, 2014.

[181] K. Hidalgo *et al.*, “Distinct physiological, biochemical and morphometric adjustments in the malaria vectors *Anopheles gambiae* and *A . coluzzii* as means to survive dry season conditions in Burkina Faso,” *J. Exp. Biol.*, vol. 221, no. 6, p. jeb174433, Mar. 2018.

[182] K. R. Reidenbach, C. Cheng, F. Liu, C. Liu, N. J. Besansky, and Z. Syed, “Cuticular differences associated with aridity acclimation in African malaria vectors carrying alternative arrangements of inversion 2La,” *Parasit. Vectors*, vol. 7, no. 1, p. 176, Apr. 2014.

[183] K. J. Gellatly, K. S. Yoon, J. J. Doherty, W. Sun, B. R. Pittendrigh, and J. M. Clark, “RNAi

validation of resistance genes and their interactions in the highly DDT-resistant 91-R strain of Drosophila melanogaster," *Pestic. Biochem. Physiol.*, vol. 121, pp. 107–115, Jun. 2015.

[184] J. P. Strycharz *et al.*, "Resistance in the highly DDT-resistant 91-R strain of Drosophila melanogaster involves decreased penetration, increased metabolism, and direct excretion," *Pestic. Biochem. Physiol.*, vol. 107, no. 2, pp. 207–217, Oct. 2013.

[185] J. H. Kim *et al.*, "Identification and interaction of multiple genes resulting in DDT resistance in the 91-R strain of Drosophila melanogaster by RNAi approaches," *Pestic. Biochem. Physiol.*, vol. 151, pp. 90–99, Oct. 2018.

[186] M. Kefi, V. Balabanidou, V. Douris, G. Lycett, R. Feyereisen, and J. Vontas, "Two functionally distinct CYP4G genes of Anopheles gambiae contribute to cuticular hydrocarbon biosynthesis," *Insect Biochem. Mol. Biol.*, vol. 110, pp. 52–59, Jul. 2019.

[187] E. L. Arrese and J. L. Soulages, "Insect Fat Body: Energy, Metabolism, and Regulation," *Annu. Rev. Entomol.*, vol. 55, no. 1, pp. 207–225, Jan. 2010.

[188] A. Lynd *et al.*, "Development of a functional genetic tool for Anopheles gambiae oenocyte characterisation: application to cuticular hydrocarbon synthesis," *bioRxiv*, p. 742619, Aug. 2019.

[189] O. Beckonert *et al.*, "Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts," *Nat. Protoc.*, vol. 2, no. 11, pp. 2692–2703, Oct. 2007.

[190] A. L. Van Geet, "Calibration of methanol nuclear magnetic resonance thermometer at low temperature," *Anal. Chem.*, vol. 42, no. 6, pp. 679–680, May 1970.

[191] L. W. Sumner *et al.*, "Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)," *Metabolomics*, vol. 3, no. 3, pp. 211–221, Sep. 2007.

[192] R Core Development Team, "R: a language and environment for statistical computing, 3.2.1," *Doc. Free. available internet http//www. r-project. org*, 2015.

[193] R. C. Team and R Core Team, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2014.

[194] C. O. Wilke, "cowplot," 2014. [Online]. Available: https://cran.r-project.org/web/packages/cowplot/index.html. [Accessed: 21-May-2019].

[195] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[196] F. Rohart, B. Gautier, A. Singh, and K.-A. Lê Cao, "mixOmics: An R package for 'omics feature selection and multiple data integration," *PLOS Comput. Biol.*, vol. 13, no. 11,

p. e1005752, Nov. 2017.

[197] J. Hao, W. Astle, M. De Iorio, and T. M. D. Ebbels, “BATMAN--an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model,” *Bioinformatics*, vol. 28, no. 15, pp. 2088–2090, Aug. 2012.

[198] H. Wickham, “Reshaping Data with the **reshape** Package,” *J. Stat. Softw.*, vol. 21, no. 12, pp. 1–20, Nov. 2007.

[199] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, “The sva package for removing batch effects and other unwanted variation in high-throughput experiments.,” *Bioinformatics*, vol. 28, no. 6, pp. 882–3, Mar. 2012.

[200] Frank Dieterle *et al.*, “Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics,” *Anal. Chem.*, vol. 78, pp. 4281–4290, 2006.

[201] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[202] N. Akarachantachote, S. Chadcham, K. Saithanu, N. Akarachantachote, S. Chadcham, and K. Saithanu, “P A CUTOFF THRESHOLD OF VARIABLE IMPORTANCE IN PROJECTION FOR VARIABLE SELECTION,” *Int. J. Pure Appl. Math.*, vol. 94, no. 3, pp. 307–322, 2014.

[203] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *J. R. Stat. Soc.*, vol. 57, no. 1, pp. 289–300, 1995.

[204] B. L. Welch, “The generalisation of student’s problems when several different population variances are involved.,” *Biometrika*, vol. 34, no. 1–2, pp. 28–35, Jan. 1947.

[205] Student, “The Probable Error of a Mean,” *Biometrika*, vol. 6, no. 1, p. 1, 1908.

[206] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “KEGG as a reference resource for gene and protein annotation,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D457–D462, Jan. 2016.

[207] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.,” *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, Jan. 2009.

[208] P. K. Mishra *et al.*, “Biochemical studies of comparative haemolymph constituents in fourth instar larvae of Daba trivoltine ecorace of topical tasar silkworm Antheraea mylitta D,” 2010.

[209] J. Lozano and X. Belles, “Conserved repressive function of Krüppel homolog 1 on insect metamorphosis in hemimetabolous and holometabolous species,” *Sci. Rep.*, vol. 1, 2011.

[210] D. J. Candy and B. A. Kilby, *Insect Biochemistry and Function*. Boston, MA: Springer US, 1975.

[211] G. J. Lycett *et al.*, "Anopheles gambiae P450 reductase is highly expressed in oenocytes and in vivo knockdown increases permethrin susceptibility," *Insect Mol. Biol.*, vol. 15, no. 3, pp. 321–327, Jun. 2006.

[212] A. Lynd, V. Balabanidou, J. Vontas, and G. J. Lycett, "Development of a functional genetic tool for Anopheles gambiae oenocyte characterisation: appliction to cuticular hydrocarbon synthesis," *Manuscr. Submitt. Publ.*, 2019.

[213] M. Namountougou *et al.*, "Multiple Insecticide Resistance in Anopheles gambiae s.l. Populations from Burkina Faso, West Africa," *PLoS One*, vol. 7, no. 11, p. e48412, Nov. 2012.

[214] V. Ingham, S. Wagstaff, and H. Ranson, "Transcriptomic meta-signatures identified in Anopheles gambiae populations reveal previously undetected insecticide resistance mechanisms," *Nat. Commun.*, vol. 9, no. 1, p. 5282, Dec. 2018.

[215] L. J. Thorat, S. M. Gaikwad, and B. B. Nath, "Trehalose as an indicator of desiccation stress in Drosophila melanogaster larvae: A potential marker of anhydrobiosis," *Biochem. Biophys. Res. Commun.*, vol. 419, no. 4, pp. 638–642, Mar. 2012.

[216] G. Iturriaga, R. Suárez, and B. Nova-Franco, "Trehalose Metabolism: From Osmoprotection to Signaling," *Int. J. Mol. Sci.*, vol. 10, no. 9, p. 3793, Sep. 2009.

[217] H.-J. Lee, Y.-S. Yoon, and S.-J. Lee, "Mechanism of neuroprotection by trehalose: controversy surrounding autophagy induction.," *Cell Death Dis.*, vol. 9, no. 7, p. 712, Jun. 2018.

[218] Y. Kanamori *et al.*, "The trehalose transporter 1 gene sequence is conserved in insects and encodes proteins with different kinetic properties involved in trehalose import into peripheral tissues," *Insect Biochem. Mol. Biol.*, vol. 40, no. 1, pp. 30–37, Jan. 2010.

[219] T. Kikawada *et al.*, "Trehalose transporter 1, a facilitated and high-capacity trehalose transporter, allows exogenous trehalose uptake into cells," vol. 104, no. 28, pp. 11585–11590, Jul. 2007.

[220] L. L. Drake, S. D. Rodriguez, and I. A. Hansen, "Functional characterization of aquaporins and aquaglyceroporins of the yellow fever mosquito, Aedes aegypti," *Sci. Rep.*, vol. 5, no. 1, p. 7795, Jul. 2015.

[221] S. Kikuta, Y. Hagiwara-Komoda, H. Noda, and T. Kikawada, "A Novel Member of the Trehalose Transporter Family Functions as an H+-Dependent Trehalose Transporter in the Reabsorption of Trehalose in Malpighian Tubules," *Front. Physiol.*, vol. 3, p. 290, Jul. 2012.

[222] E. Shukla, L. J. Thorat, B. B. Nath, and S. M. Gaikwad, “Insect trehalase: Physiological significance and potential applications,” *Glycobiology*, vol. 25, no. 4, pp. 357–367, 2015.

[223] R. G. H. Downer and J. R. Matthews, “Trehalase activity of serum and whole haemolymph in the American cockroach, *Periplaneta americana* L.,” *Can. J. Zool.*, vol. 56, no. 10, pp. 2217–2219, Oct. 1978.

[224] J. Xu, Z. Sheng, and S. R. Palli, “Juvenile Hormone and Insulin Regulate Trehalose Homeostasis in the Red Flour Beetle, Tribolium castaneum,” *PLoS Genet.*, vol. 9, no. 6, p. e1003535, Jun. 2013.

[225] Q. Chen, “Role of trehalose phosphate synthase and trehalose during hypoxia: from flies to mammals,” *J. Exp. Biol.*, vol. 207, no. 18, pp. 3125–3129, Aug. 2004.

[226] M. J. Paul, M. Oszvald, C. Jesus, C. Rajulu, and C. A. Griffiths, “Increasing crop yield and resilience with trehalose 6-phosphate: targeting a feast–famine mechanism in cereals for better source–sink optimization,” *J. Exp. Bot.*, vol. 68, no. 16, pp. 4455–4462, Jul. 2017.

[227] J. I. Vílchez, C. García-Fontana, D. Román-Naranjo, J. González-López, and M. Manzanera, “Plant Drought Tolerance Enhancement by Trehalose Production of Desiccation-Tolerant Microorganisms.,” *Front. Microbiol.*, vol. 7, p. 1577, 2016.

[228] N. Asano, M. Takeuchi, Y. Kameda, K. Matsui, and Y. Kono, “Trehalase inhibitors, validoxylamine A and related compounds as insecticides.,” *J. Antibiot. (Tokyo).*, vol. 43, no. 6, pp. 722–726, 1990.

[229] L.-Q. Jin, Y.-P. Xue, Y.-G. Zheng, and Y.-C. Shen, “Production of trehalase inhibitor validoxylamine A using acid-catalyzed hydrolysis of validamycin A,” *Catal. Commun.*, vol. 7, no. 3, pp. 157–161, Mar. 2006.

[230] R. D. Hall, “Plant metabolomics: from holistic hope, to hype, to hot topic,” *New Phytol.*, vol. 169, no. 3, pp. 453–468, Feb. 2006.

[231] C. B. Prasannan, D. Jaiswal, R. Davis, and P. P. Wangikar, “An improved method for extraction of polar and charged metabolites from cyanobacteria,” *PLoS One*, vol. 13, no. 10, 2018.

[232] S. C. Sapcariu, T. Kanashova, D. Weindl, J. Ghelfi, G. Dittmar, and K. Hiller, “Simultaneous extraction of proteins and metabolites from cells in culture,” *MethodsX*, vol. 1, pp. 74–80, 2014.

[233] P. Masson, A. C. Alves, T. M. D. Ebbels, J. K. Nicholson, and E. J. Want, “Optimization and Evaluation of Metabolite Extraction Protocols for Untargeted Metabolic Profiling of Liver Samples by UPLC-MS,” *Anal. Chem.*, vol. 82, no. 18, pp. 7779–7786, Sep. 2010.

[234] M. Coen, E. M. Lenz, J. K. Nicholson, I. D. Wilson, F. Pognan, and J. C. Lindon, "An integrated metabonomic investigation of acetaminophen toxicity in the mouse using NMR spectroscopy," *Chem. Res. Toxicol.*, vol. 16, no. 3, pp. 295–303, Mar. 2003.

[235] Y. L. Ching *et al.*, "Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics," *Metabolomics*, vol. 3, no. 1, pp. 55–67, 2007.

[236] G. D. Stentiford *et al.*, "Liver Tumors in Wild Flatfish: A Histopathological, Proteomic, and Metabolomic Study," *Omi. A J. Integr. Biol.*, vol. 9, no. 3, pp. 281–299, Sep. 2005.

[237] M. Vinaixa, S. Samino, I. Saez, J. Duran, J. J. Guinovart, and O. Yanes, "A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data," *Metabolites*, vol. 2, pp. 775–795, 2012.

[238] E. Anderssen, K. Dyrstad, F. Westad, and H. Martens, "Reducing over-optimism in variable selection by cross-model validation," *Chemom. Intell. Lab. Syst.*, vol. 84, no. 1–2, pp. 69–74, Dec. 2006.

[239] C. A. Westerhuis *et al.*, "Assessment of PLSDA cross validation," *Metabolomics*, vol. 4, no. 1, pp. 81–89, Mar. 2007.

[240] P. S. Gromski *et al.*, "A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding," *Anal. Chim. Acta*, vol. 879, pp. 10–23, Jun. 2015.

[241] D. S. Wishart *et al.*, "HMDB 3.0-The Human Metabolome Database in 2013," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D801-7, Jan. 2013.

[242] S. A. Goldansaz, A. C. Guo, T. Sajed, M. A. Steele, G. S. Plastow, and D. S. Wishart, "Livestock metabolomics and the livestock metabolome: A systematic review," *PLoS One*, vol. 12, no. 5, p. e0177675, May 2017.

8 Appendices

Appendix 1: Pattern file for *Anopheles* and *Aedes* datasets. WT and knock-down *Anohpeles* datasets use the same pattern file. Also, pupa and adult stages use the same pattern file of the same group. UNID: unidentified; NA: Not applicable.

Right [ppm]	Left [ppm]	Anopheles bins	Aedes bins
0.8769	0.9021	UNID.1	UNID.1
0.9316	0.9349	Isoleucine.2	Isoleucine.2
0.9349	0.9386	UNID.3	UNID.3
0.9412	0.9469	Isoleucine.4	Isoleucine.4
0.9521	0.9550	UNID.5	UNID.5
0.9526	0.9658	UNID.6	UNID.6
0.9658	0.9824	UNID.7	UNID.7
0.9883	1.0032	Valine.8	Valine.8
1.0077	1.0238	Isoleucine.9	Isoleucine.9
1.0374	1.0437	Valine.10	Valine.10
1.0437	1.0464	UNID.11	UNID.11
1.0464	1.0495	Propionate.12	Propionate.12
1.0495	1.0540	Valine.13	Valine.13
1.0540	1.0561	UNID.14	UNID.14
1.0579	1.0600	Propionate.15	Propionate.15
1.0600	1.0622	UNID.16	UNID.16
1.0686	1.0726	Propionate.17	Propionate.17
1.0838	1.1079	UNID.18	UNID.18
1.1378	1.1396	UNID.20	UNID.20
1.1396	1.1424	UNID.21	UNID.21
1.1460	1.1480	UNID.22	UNID.22
1.1480	1.1498	UNID.23	UNID.23
1.1498	1.1513	UNID.24	UNID.24
1.1582	1.1602	UNID.25	UNID.25
1.1602	1.1643	UNID.26	UNID.26
1.1643	1.1660	UNID.27	UNID.27
1.1660	1.1706	UNID.28	UNID.28
1.1706	1.1733	UNID.29	UNID.29
1.1733	1.1793	UNID.30	UNID.30
1.1793	1.1823	UNID.31	UNID.31
1.1823	1.1903	UNID.32	UNID.32
1.1903	1.1931	UNID.33	UNID.33
1.1931	1.1960	UNID.34	UNID.34
1.1960	1.2006	UNID.35	UNID.35
1.2006	1.2050	UNID.36	UNID.36
1.2117	1.2141	UNID.37	UNID.37
1.2141	1.2207	UNID.38	UNID.38
1.2207	1.2226	UNID.39	UNID.39
1.2226	1.2249	UNID.40	UNID.40
1.2249	1.2278	UNID.41	UNID.41
1.2278	1.2315	UNID.42	UNID.42
1.2315	1.2362	UNID.43	UNID.43
1.2362	1.2409	UNID.44	UNID.44
1.2409	1.2446	UNID.45	UNID.45
1.2446	1.2477	UNID.46	UNID.46
1.2584	1.3239	UNID.48	UNID.48
1.3239	1.3278	Lactate.49	Lactate.49
1.3278	1.3319	Threonine.50	Theronine.50
1.3319	1.3377	Lactate.51	Lactate.51
1.3377	1.3400	Threonine.52	Theronine.52
1.4381	1.4448	UNID.53	UNID.53
1.4448	1.4536	UNID.54	UNID.54
1.4536	1.4574	UNID.55	UNID.55
1.4574	1.4609	UNID.56	UNID.56
1.4609	1.4684	UNID.57	UNID.57
1.4732	1.4925	Alanine.58	Alanine.58
1.4981	1.5369	UNID.59	UNID.59
1.6322	1.6393	UNID.60	UNID.60
1.6393	1.6432	UNID.61	UNID.61
1.6432	1.6491	UNID.62	UNID.62
1.6491	1.6529	UNID.63	UNID.63
1.6529	1.6584	UNID.64	UNID.64
1.6584	1.6626	UNID.65	UNID.65
1.6626	1.6679	UNID.66	UNID.66

1.6679	1.6731	UNID.67	UNID.67
1.6779	1.6856	UNID.68	UNID.68
1.6856	1.6914	UNID.69	UNID.69
1.6914	1.6963	UNID.70	UNID.70
1.6963	1.7026	UNID.71	UNID.71
1.7026	1.7077	UNID.72	UNID.72
1.7077	1.7109	UNID.73	UNID.73
1.7109	1.7192	UNID.74	UNID.74
1.7192	1.7275	UNID.75	UNID.75
1.7275	1.7312	UNID.76	UNID.76
1.7312	1.7364	UNID.77	UNID.77
1.7364	1.7412	UNID.78	UNID.78
1.7414	1.7463	UNID.79	UNID.79
1.7463	1.7503	UNID.80	UNID.80
1.7503	1.7536	UNID.81	UNID.81
1.7536	1.7561	UNID.82	UNID.82
1.7561	1.7603	UNID.83	UNID.83
1.7603	1.7646	UNID.84	UNID.84
1.7646	1.7698	UNID.85	UNID.85
1.7698	1.7725	UNID.86	UNID.86
1.7725	1.7763	UNID.87	UNID.87
1.7763	1.7822	UNID.88	UNID.88
1.7822	1.7877	UNID.89	UNID.89
1.7877	1.7924	UNID.90	UNID.90
1.7924	1.7985	UNID.91	UNID.91
1.8115	1.8158	NA	UNID.92
1.8201	1.8228	NA	UNID.93
1.8228	1.8261	NA	UNID.94
1.8261	1.8307	UNID.92	UNID.95
1.8307	1.8351	NA	UNID.96
1.8351	1.8379	NA	UNID.97
1.8379	1.8429	NA	UNID.98
1.8429	1.8459	UNID.93	UNID.99
1.8459	1.8504	NA	UNID.100
1.8504	1.8542	NA	UNID.101
1.8542	1.8578	NA	UNID.102
1.8578	1.8618	NA	UNID.103
1.8656	1.8691	NA	UNID.104
1.8421	1.8474	NA	UNID.105
1.8658	1.8702	UNID.94	UNID.106
1.8749	1.8821	UNID.95	UNID.107
1.8835	1.8888	UNID.96	UNID.108
1.8888	1.8916	UNID.97	UNID.109
1.8916	1.8970	UNID.98	UNID.110
1.8970	1.9017	UNID.99	UNID.111
1.9017	1.9112	UNID.100	UNID.112
1.9112	1.9168	UNID.101	UNID.113
1.9168	1.9228	Acetate.102	Acetate.114
1.9228	1.9267	UNID.103	UNID.115
1.9267	1.9312	UNID.104	UNID.116
1.9312	1.9349	UNID.105	UNID.117
1.9349	1.9423	UNID.106	UNID.118
1.9423	1.9450	UNID.107	UNID.119
1.9554	1.9617	UNID.108	UNID.120
1.9711	1.9774	UNID.109	UNID.121
1.9899	1.9977	UNID.111	UNID.123
1.9977	2.0010	UNID.112	UNID.124
2.0010	2.0055	UNID.113	UNID.125
2.0055	2.0076	UNID.114	UNID.126
2.0076	2.0099	UNID.115	UNID.127
2.0099	2.0114	UNID.116	UNID.128
2.0114	2.0161	UNID.117	UNID.129
2.0161	2.0795	Glutamate.118	Glutamate.130
2.0795	2.0833	UNID.119	UNID.131
2.0833	2.0861	UNID.120	UNID.132
2.0861	2.0879	UNID.121	UNID.133
2.0879	2.0895	UNID.122	UNID.134
2.0895	2.0912	UNID.123	UNID.135
2.0912	2.0959	UNID.124	UNID.136
2.0959	2.1048	UNID.125	UNID.137
2.1092	2.1666	Glutamate.Glutamine.126	Glutamate.Glutamine.138

2.1666	2.1688	Propionate.127	Propionate.139
2.1688	2.1773	Glutamine.128	Glutamine.140
2.1773	2.1917	Propionate.129	Propionate.141
2.1983	2.2026	Propionate.130	Propionate.142
2.2323	2.2389	Glutamine.131	Glutamine.143
2.2389	2.3040	Valine.132	Valine.144
2.3091	2.3119	UNID.133	UNID.145
2.3119	2.3169	UNID.134	UNID.146
2.3169	2.3202	Glutamate.135	Glutamate.147
2.3202	2.3235	UNID.136	UNID.148
2.3235	2.3268	UNID.137	UNID.149
2.3268	2.3654	Glutamate.138	Glutamate.150
2.3654	2.3688	Pyruvate.139	Pyruvate.151
2.3688	2.3730	Glutamate.140	Glutamate.152
2.3730	2.3786	UNID.141	UNID.153
2.3786	2.3950	Glutamate.142	Glutamate.154
2.3950	2.4009	UNID.143	UNID.155
2.4009	2.4132	Succinate.144	Succinate.156
2.4132	2.4205	UNID.145	UNID.157
2.4205	2.4242	UNID.146	UNID.158
2.4242	2.4309	UNID.147	UNID.159
2.4309	2.4345	UNID.148	UNID.160
2.4345	2.4412	UNID.149	UNID.161
2.4412	2.4453	UNID.150	UNID.162
2.4453	2.4504	UNID.151	UNID.163
2.4504	2.4525	UNID.152	UNID.164
2.4525	2.4555	UNID.153	UNID.165
2.4596	2.4624	UNID.154	UNID.166
2.4624	2.4647	UNID.155	UNID.167
2.4647	2.4682	UNID.156	UNID.168
2.4700	2.4730	UNID.157	UNID.169
2.4730	2.4781	UNID.158	UNID.170
2.4832	2.4866	UNID.159	UNID.171
2.4866	2.4899	UNID.160	UNID.172
2.4955	2.5011	UNID.161	UNID.173
2.5081	2.5128	UNID.162	UNID.174
2.5146	2.5181	UNID.163	UNID.175
2.5181	2.5236	UNID.164	UNID.176
2.5236	2.5574	UNID.165	UNID.177
2.5574	2.5625	UNID.166	UNID.178
2.5643	2.5675	UNID.167	UNID.179
2.5675	2.5713	UNID.168	UNID.180
2.6343	2.6388	UNID.169	UNID.181
2.6458	2.6495	UNID.170	UNID.182
2.6495	2.6570	UNID.171	UNID.183
2.6570	2.6621	UNID.172	UNID.184
2.6621	2.6674	UNID.173	UNID.185
2.6674	2.6791	UNID.174	UNID.186
2.6791	2.6841	UNID.175	UNID.187
2.6841	2.6914	UNID.176	UNID.188
2.6985	2.7092	UNID.177	UNID.189
2.7092	2.7105	UNID.178	UNID.190
2.7105	2.7131	UNID.179	UNID.191
2.7359	2.7399	UNID.180	UNID.192
2.7399	2.7420	UNID.181	UNID.193
2.7501	2.7593	UNID.182	UNID.194
2.7593	2.7619	UNID.183	UNID.195
2.7876	2.7936	NA	UNID.196
2.7972	2.8052	NA	UNID.197
2.8002	2.8050	UNID.184	UNID.198
2.8050	2.8101	UNID.185	UNID.199
2.8167	2.8201	UNID.186	UNID.200
2.8258	2.8304	UNID.187	UNID.201
2.8304	2.8334	UNID.188	UNID.202
2.8334	2.8362	UNID.189	UNID.203
2.8441	2.8481	UNID.190	UNID.204
2.8481	2.8519	UNID.191	UNID.205
2.8558	2.8602	UNID.192	UNID.206
2.8675	2.8852	UNID.193	UNID.207
2.9178	2.9227	UNID.194	UNID.208
2.9227	2.9269	UNID.195	UNID.209

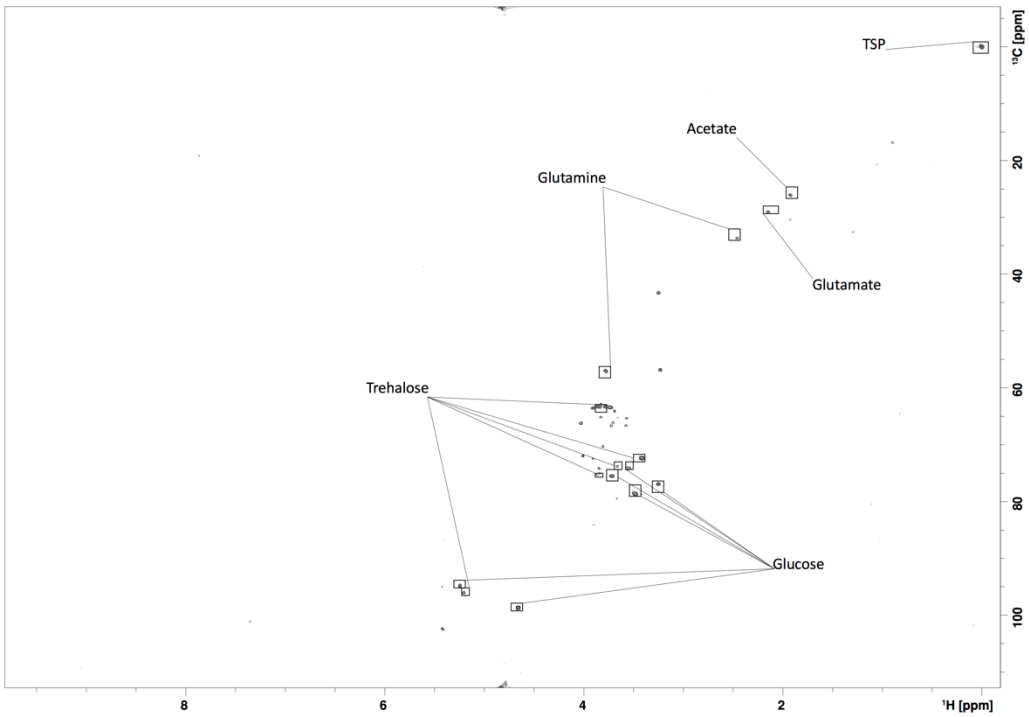
2.9382	2.9496	UNID.196	UNID.210
2.9636	2.9731	UNID.197	UNID.211
2.9868	2.9922	UNID.198	UNID.212
2.9979	3.0024	UNID.199	UNID.213
3.0024	3.0068	UNID.200	UNID.214
3.0068	3.0131	UNID.201	UNID.215
3.0131	3.0188	UNID.202	UNID.216
3.0188	3.0235	UNID.203	UNID.217
3.0235	3.0343	UNID.204	UNID.218
3.0343	3.0411	UNID.205	UNID.219
3.0411	3.0434	UNID.206	UNID.220
3.0434	3.0487	Tyrosine.207	Tyrosine.221
3.0487	3.0510	UNID.208	UNID.222
3.0510	3.0545	UNID.209	UNID.223
3.0545	3.0601	Tyrosine.210	Tyrosine.224
3.0601	3.0644	UNID.211	UNID.225
3.0644	3.0723	Tyrosine.212	Tyrosine.226
3.0723	3.0760	UNID.213	UNID.227
3.0760	3.0823	Tyrosine.214	Tyrosine.228
3.0823	3.0858	UNID.215	UNID.229
3.0885	3.0934	UNID.216	UNID.230
3.1007	3.1050	UNID.217	UNID.231
3.1077	3.1103	UNID.218	UNID.232
3.1175	3.1204	UNID.219	UNID.233
3.1204	3.1250	UNID.220	UNID.234
3.1250	3.1284	UNID.221	UNID.235
3.1284	3.1332	UNID.222	UNID.236
3.1351	3.1378	UNID.223	UNID.237
3.1378	3.1556	UNID.224	UNID.238
3.1584	3.1779	UNID.225	UNID.239
3.1779	3.1851	UNID.226	UNID.240
3.1851	3.2015	Tyrosine.227	Tyrosine.241
3.2015	3.2050	UNID.228	UNID.242
3.2050	3.2093	UNID.229	UNID.243
3.2093	3.2141	Tyrosine.230	Tyrosine.244
3.2141	3.2169	UNID.231	UNID.245
3.2169	3.2210	Tyrosine.232	Tyrosine.246
3.2210	3.2284	UNID.233	UNID.247
3.2284	3.2329	UNID.234	UNID.248
3.2329	3.2359	UNID.235	UNID.249
3.2359	3.2540	Glucose.236	Glucose.250
3.2540	3.2563	UNID.237	UNID.251
3.2563	3.2603	UNID.238	UNID.252
3.2603	3.2655	Glucose.239	Glucose.253
3.2655	3.2700	UNID.240	UNID.254
3.2700	3.2734	UNID.241	UNID.255
3.2734	3.2758	UNID.242	UNID.256
3.2758	3.2789	UNID.243	UNID.257
3.2789	3.2810	UNID.244	UNID.258
3.2810	3.2866	UNID.245	UNID.259
3.2866	3.2924	UNID.246	UNID.260
3.2924	3.2983	UNID.247	UNID.261
3.2983	3.3021	Tryptophan.248	Tryptophan.262
3.3021	3.3053	UNID.249	UNID.263
3.3053	3.3086	UNID.250	UNID.264
3.3086	3.3137	Tryptophan.251	Tryptophan.265
3.3191	3.3231	Tryptophan.252	Tryptophan.266
3.3231	3.3265	UNID.253	UNID.267
3.3265	3.3294	UNID.254	UNID.268
3.3294	3.3376	Tryptophan.255	Tryptophan.269
3.3376	3.3429	UNID.256	UNID.270
3.3429	3.3475	UNID.257	UNID.271
3.3475	3.3551	UNID.258	UNID.272
3.3551	3.3595	UNID.259	UNID.273
3.3595	3.3638	Methanol.260	Methanol.274
3.3638	3.3702	UNID.261	UNID.275
3.3702	3.3865	UNID.262	UNID.276
3.3865	3.3908	UNID.263	UNID.277
3.3908	3.3962	Glucose.264	Glucose.278
3.3962	3.4021	UNID.265	UNID.279
3.4021	3.4044	UNID.266	UNID.280

3.4044	3.4086	Glucose.267	Glucose.281
3.4086	3.4131	UNID.268	UNID.282
3.4131	3.4182	UNID.269	UNID.283
3.4182	3.4235	Glucose.270	Glucose.284
3.4235	3.4272	UNID.271	UNID.285
3.4272	3.4287	UNID.272	UNID.286
3.4287	3.4313	UNID.273	UNID.287
3.4313	3.4385	Glucose.274	Glucose.288
3.4385	3.4449	Trehalose.275	Trehalose.289
3.4449	3.4516	Glucose.276	Glucose.290
3.4516	3.4560	Trehalose.277	Trehalose.291
3.4560	3.4599	Glucose.Trehalose.278	Glucose.Trehalose.292
3.4599	3.4678	Glucose.279	Glucose.293
3.4678	3.4714	Glucose.Trehalose.280	Glucose.Trehalose.294
3.4714	3.4770	Glucose.281	Glucose.295
3.4770	3.4828	Glucose.Tryptophan.282	Glucose.Tryptophan.296
3.4828	3.4850	Glucose.283	Glucose.297
3.4850	3.4898	Tryptophan.284	Tryptophan.298
3.4898	3.4984	Glucose.285	Glucose.299
3.4984	3.5071	Tryptophan.286	Tryptophan.300
3.5071	3.5102	Glucose.287	Glucose.301
3.5162	3.5217	UNID.288	UNID.302
3.5262	3.5312	Glucose.289	Glucose.303
3.5312	3.5335	UNID.290	UNID.304
3.5335	3.5375	Glucose.291	Glucose.305
3.5407	3.5520	Glucose.292	Glucose.306
3.5580	3.5600	UNID.293	UNID.307
3.5600	3.5643	UNID.294	UNID.308
3.5643	3.5706	Glycine.295	Glycine.309
3.5706	3.5750	UNID.296	UNID.310
3.5750	3.5779	UNID.297	UNID.311
3.5779	3.5838	UNID.298	UNID.312
3.5838	3.5932	Threonine.299	Theronine.313
3.5932	3.5968	UNID.300	UNID.314
3.5968	3.6017	UNID.301	UNID.315
3.6017	3.6045	UNID.302	UNID.316
3.6045	3.6098	UNID.303	UNID.317
3.6098	3.6188	Valine.304	Valine.318
3.6188	3.6230	UNID.305	UNID.319
3.6245	3.6280	UNID.306	UNID.320
3.6280	3.6302	UNID.307	UNID.321
3.6302	3.6333	UNID.308	UNID.322
3.6333	3.6371	UNID.309	UNID.323
3.6371	3.6418	UNID.310	UNID.324
3.6418	3.6505	Trehalose.311	Trehalose.325
3.6505	3.6559	UNID.312	UNID.326
3.6559	3.6588	Trehalose.313	Trehalose.327
3.6588	3.6619	UNID.314	UNID.328
3.6619	3.6648	Trehalose.315	Trehalose.329
3.6648	3.6693	UNID.316	UNID.330
3.6693	3.6751	UNID.317	UNID.331
3.6751	3.6804	UNID.318	UNID.332
3.6804	3.6835	UNID.319	UNID.333
3.6835	3.6874	UNID.320	UNID.334
3.6874	3.6940	UNID.321	UNID.335
3.6940	3.7024	UNID.322	UNID.336
3.7024	3.7083	Glucose.323	Glucose.337
3.7083	3.7107	UNID.324	UNID.338
3.7107	3.7128	UNID.325	UNID.339
3.7128	3.7164	UNID.326	UNID.340
3.7164	3.7211	Glucose.327	Glucose.341
3.7211	3.7268	UNID.328	UNID.342
3.7268	3.7457	Glucose.329	Glucose.343
3.7457	3.7506	UNID.330	UNID.344
3.7506	3.7607	Glutamate.331	Glutamate.345
3.7607	3.7636	Glucose.Glutamate.332	Glucose.Glutamate.346
3.7636	3.7688	Trehalose.Glutamine.333	Trehalose.Glutamine.347
3.7688	3.7716	Glucose.Alanine.Glutamate.334	Glucose.Alanine.Glutamate.348
3.7716	3.7760	Trehalose.Glutamine.335	Trehalose.Glutamine.349
3.7760	3.7792	Glutamate.336	Glutamate.350
3.7792	3.7814	Glucose.Alanine.337	Glucose.Alanine.351

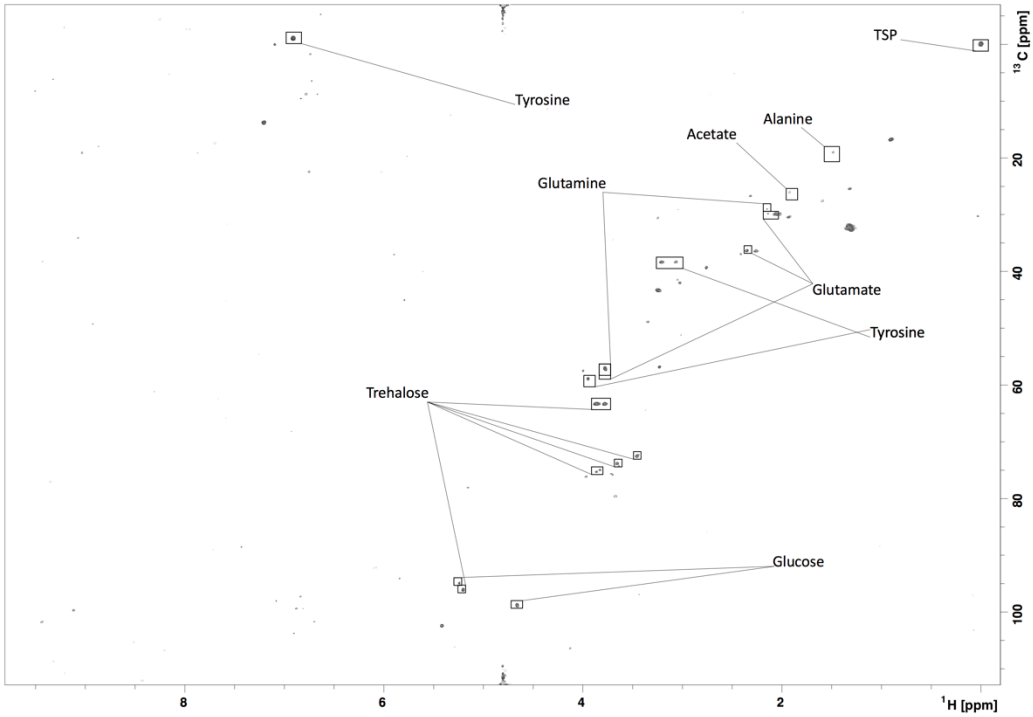
3.7814	3.7876	Glucose.338	Glucose.352
3.7876	3.7898	UNID.339	UNID.353
3.7898	3.7958	Trehalose.Alanine.340	Trehalose.Alanine.354
3.7958	3.8000	UNID.341	UNID.355
3.8000	3.8038	Alanine.342	Alanine.356
3.8126	3.8157	UNID.343	UNID.357
3.8157	3.8198	Trehalose.344	Trehalose.358
3.8198	3.8233	Trehalose.345	Trehalose.359
3.8233	3.8451	Glucose.Trehalose.346	Glucose.Trehalose.360
3.8451	3.8486	Glucose.347	Glucose.361
3.8486	3.8562	Trehalose.348	Trehalose.362
3.8562	3.8663	Glucose.Trehalose.349	Glucose.Trehalose.363
3.8663	3.8698	UNID.350	UNID.364
3.8698	3.8791	Trehalose.351	Trehalose.365
3.8825	3.8850	UNID.352	UNID.366
3.8850	3.8881	UNID.353	UNID.367
3.8881	3.8946	Glucose.354	Glucose.368
3.9004	3.9049	UNID.355	UNID.369
3.9049	3.9136	Glucose.356	Glucose.370
3.9185	3.9228	UNID.357	UNID.371
3.9228	3.9272	UNID.358	UNID.372
3.9272	3.9301	UNID.359	UNID.373
3.9330	3.9559	Tyrosine.360	Tyrosine.374
3.9604	3.9676	UNID.361	UNID.375
3.9747	3.9814	UNID.362	UNID.376
3.9814	3.9836	UNID.363	UNID.377
3.9836	3.9863	UNID.364	UNID.378
3.9863	3.9892	UNID.365	UNID.379
3.9892	3.9944	UNID.366	UNID.380
3.9944	3.9964	UNID.367	UNID.381
3.9964	3.9995	UNID.368	UNID.382
3.9995	4.0016	UNID.369	UNID.383
4.0016	4.0047	UNID.370	UNID.384
4.0047	4.0070	UNID.371	UNID.385
4.0070	4.0101	UNID.372	UNID.386
4.0101	4.0134	UNID.373	UNID.387
4.0134	4.0152	UNID.374	UNID.388
4.0152	4.0173	UNID.375	UNID.389
4.0173	4.0209	UNID.376	UNID.390
4.0209	4.0258	UNID.377	UNID.391
4.0258	4.0307	UNID.378	UNID.392
4.0307	4.0344	UNID.379	UNID.393
4.0344	4.0388	UNID.380	UNID.394
4.0388	4.0409	UNID.381	UNID.395
4.0409	4.0431	UNID.382	UNID.396
4.0537	4.0766	Tryptophan.383	Tryptophan.397
4.0961	4.1219	Lactate.384	Lactate.398
4.1219	4.1268	UNID.385	UNID.399
4.1268	4.1296	Lactate.386	Lactate.400
4.1296	4.1327	UNID.387	UNID.401
4.1327	4.1354	UNID.388	UNID.402
4.1354	4.1372	UNID.389	UNID.403
4.1372	4.1411	UNID.390	UNID.404
4.1450	4.1505	UNID.391	UNID.405
4.1566	4.1609	UNID.392	UNID.406
4.1609	4.1647	UNID.393	UNID.407
4.1831	4.1916	UNID.394	UNID.408
4.1916	4.2017	UNID.395	UNID.409
4.2237	4.2303	UNID.396	UNID.410
4.2303	4.2356	UNID.397	UNID.411
4.2356	4.2771	Threonine.398	Theronine.412
4.2771	4.2830	UNID.399	UNID.413
4.2830	4.2890	UNID.400	UNID.414
4.2890	4.2928	UNID.401	UNID.415
4.2928	4.2954	UNID.402	UNID.416
4.2954	4.2987	UNID.403	UNID.417
4.3062	4.3099	UNID.404	UNID.418
4.3099	4.3122	UNID.405	UNID.419
4.3122	4.3148	UNID.406	UNID.420
4.3260	4.3321	UNID.407	UNID.421
4.3499	4.3560	UNID.408	UNID.422

4.3560	4.3637	UNID.409	UNID.423
4.3637	4.3716	UNID.410	UNID.424
4.3716	4.3787	UNID.411	UNID.425
4.3787	4.3854	UNID.412	UNID.426
4.4352	4.4436	UNID.413	UNID.427
4.4436	4.4523	UNID.414	UNID.428
4.4584	4.4722	UNID.415	UNID.429
4.4722	4.4851	UNID.416	UNID.430
4.5055	4.5147	UNID.417	UNID.431
4.5147	4.5232	UNID.418	UNID.432
4.5622	4.5654	UNID.419	UNID.433
4.6365	4.6577	Glucose.420	Glucose.434
4.6577	4.6643	UNID.421	UNID.435
4.6643	4.6720	UNID.422	UNID.436
5.1837	5.1925	UNID.423	UNID.437
5.1925	5.2087	Trehalose.424	Trehalose.438
5.2326	5.2504	Glucose.425	Glucose.439
5.3025	5.3843	UNID.426	UNID.440
5.3843	5.3956	UNID.427	UNID.441
5.3956	5.4019	UNID.428	UNID.442
5.4019	5.4078	UNID.429	UNID.443
5.4078	5.4138	UNID.430	UNID.444
5.4138	5.4198	UNID.431	UNID.445
5.4198	5.4294	UNID.432	UNID.446
5.7048	5.7140	UNID.433	UNID.447
5.8961	5.9048	UNID.434	UNID.448
5.9048	5.9146	UNID.435	UNID.449
5.9146	5.9225	UNID.436	UNID.450
5.9225	5.9307	UNID.437	UNID.451
5.9661	5.9768	UNID.438	UNID.452
5.9768	5.9840	UNID.439	UNID.453
5.9840	5.9899	UNID.440	UNID.454
5.9899	5.9970	UNID.441	UNID.455
6.0355	6.0438	UNID.442	UNID.456
6.0438	6.0527	UNID.443	UNID.457
6.0858	6.0931	UNID.444	UNID.458
6.0931	6.1064	UNID.445	UNID.459
6.1064	6.1178	UNID.446	UNID.460
6.1401	6.1468	UNID.447	UNID.461
6.1468	6.1555	UNID.448	UNID.462
6.1555	6.1644	UNID.449	UNID.463
6.5156	6.5279	Fumarate.450	Fumarate.464
6.6968	6.7295	UNID.451	UNID.465
6.8891	6.9202	Tyrosine.452	Tyrosine.466
6.9719	6.9776	NA	UNID.467
6.9856	6.9901	NA	UNID.468
6.9956	7.0048	UNID.453	UNID.469
7.0434	7.0526	UNID.454	UNID.470
7.0526	7.0629	UNID.455	UNID.471
7.0832	7.1003	UNID.456	UNID.472
7.1771	7.1898	Tyrosine.457	Tyrosine.473
7.1898	7.1989	Tyrosine.Tryptophan.458	Tyrosine.Tryptophan.474
7.1989	7.2028	Tyrosine.459	Tyrosine.475
7.2028	7.2080	Tyrosine.Tryptophan.460	Tyrosine.Tryptophan.476
7.2080	7.2124	Tyrosine.461	Tyrosine.477
7.2124	7.2191	Tryptophan.462	Tryptophan.478
7.2525	7.2603	NA	UNID.479
7.2740	7.3022	Tryptophan.463	Tryptophan.480
7.3062	7.3154	UNID.464	UNID.481
7.3154	7.3243	UNID.465	UNID.482
7.3243	7.3333	Tryptophan.466	Tryptophan.483
7.3333	7.3396	UNID.467	UNID.484
7.3396	7.3442	UNID.468	UNID.485
7.3758	7.3861	UNID.469	UNID.486
7.3861	7.3950	UNID.470	UNID.487
7.4169	7.4269	UNID.471	UNID.488
7.4269	7.4373	UNID.472	UNID.489
7.4373	7.4462	UNID.473	UNID.490
7.4462	7.4556	UNID.474	UNID.491
7.4556	7.4671	UNID.475	UNID.492
7.5366	7.5567	Tryptophan.476	Tryptophan.493

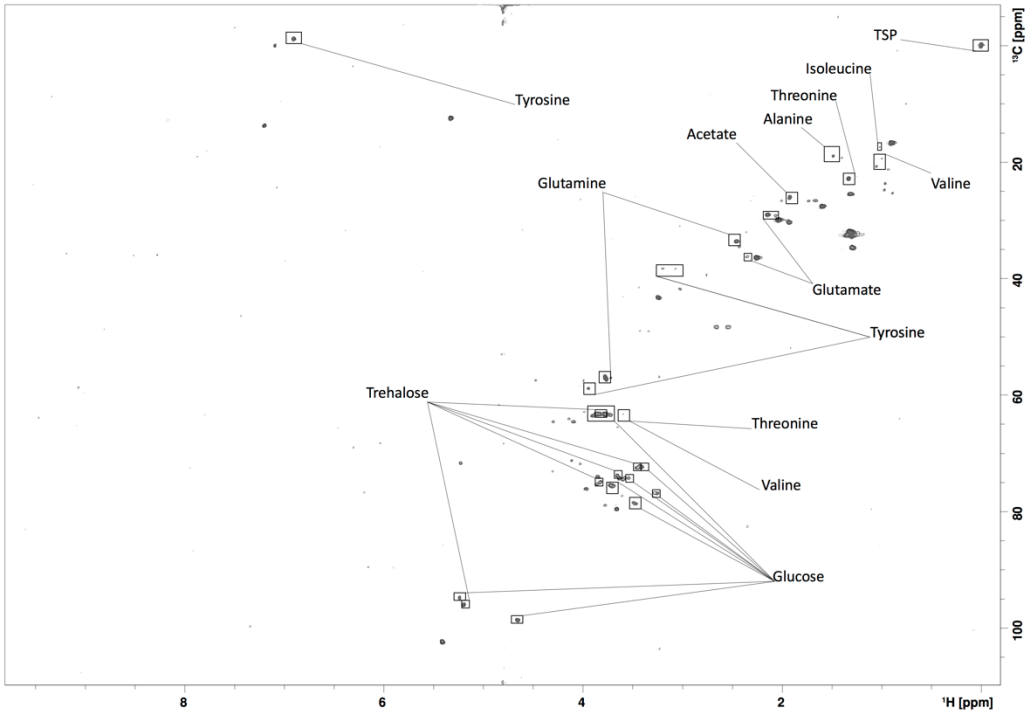
7.6911	7.7091	UNID.477	UNID.494
7.7316	7.7519	Tryptophan.478	Tryptophan.495
7.7717	7.7928	UNID.479	UNID.496
7.8557	7.8731	UNID.480	UNID.497
7.8731	7.8774	UNID.481	UNID.498
7.8837	7.8895	Xanthine.482	Xanthine.499
7.9521	7.9554	UNID.483	UNID.500
7.9554	7.9605	UNID.484	UNID.501
7.9647	7.9668	UNID.485	UNID.502
7.9668	7.9722	UNID.486	UNID.503
8.0027	8.0085	UNID.487	UNID.504
8.0975	8.1157	UNID.488	UNID.505
8.1771	8.1820	UNID.489	UNID.506
8.2100	8.2142	UNID.490	UNID.507
8.2337	8.2381	UNID.491	UNID.508
8.2381	8.2446	UNID.492	UNID.509
8.2698	8.2735	Oxypurinol.493	Oxypurinol.510
8.2735	8.2781	UNID.494	UNID.511
8.4316	8.4361	UNID.496	UNID.513
8.4521	8.4658	Formate.497	Formate.514
8.5379	8.5478	UNID.498	UNID.515
8.5826	8.5869	UNID.499	UNID.516
8.6076	8.6186	UNID.500	UNID.517



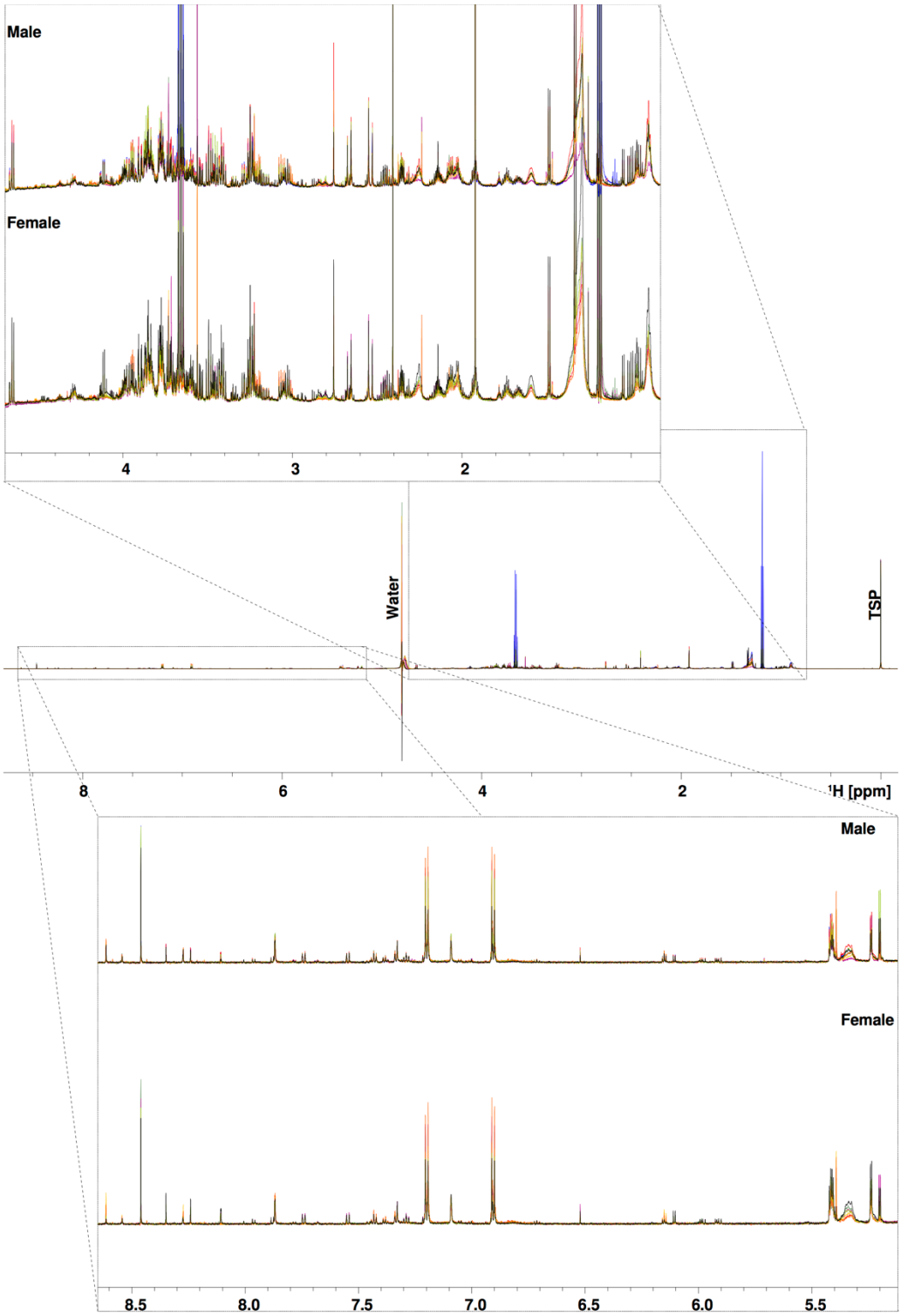
Appendix 2: Metabolite assignment on a ^{1}H-^{13}C HSQC spectrum from a representative *An. gambiae* knock-down sample.



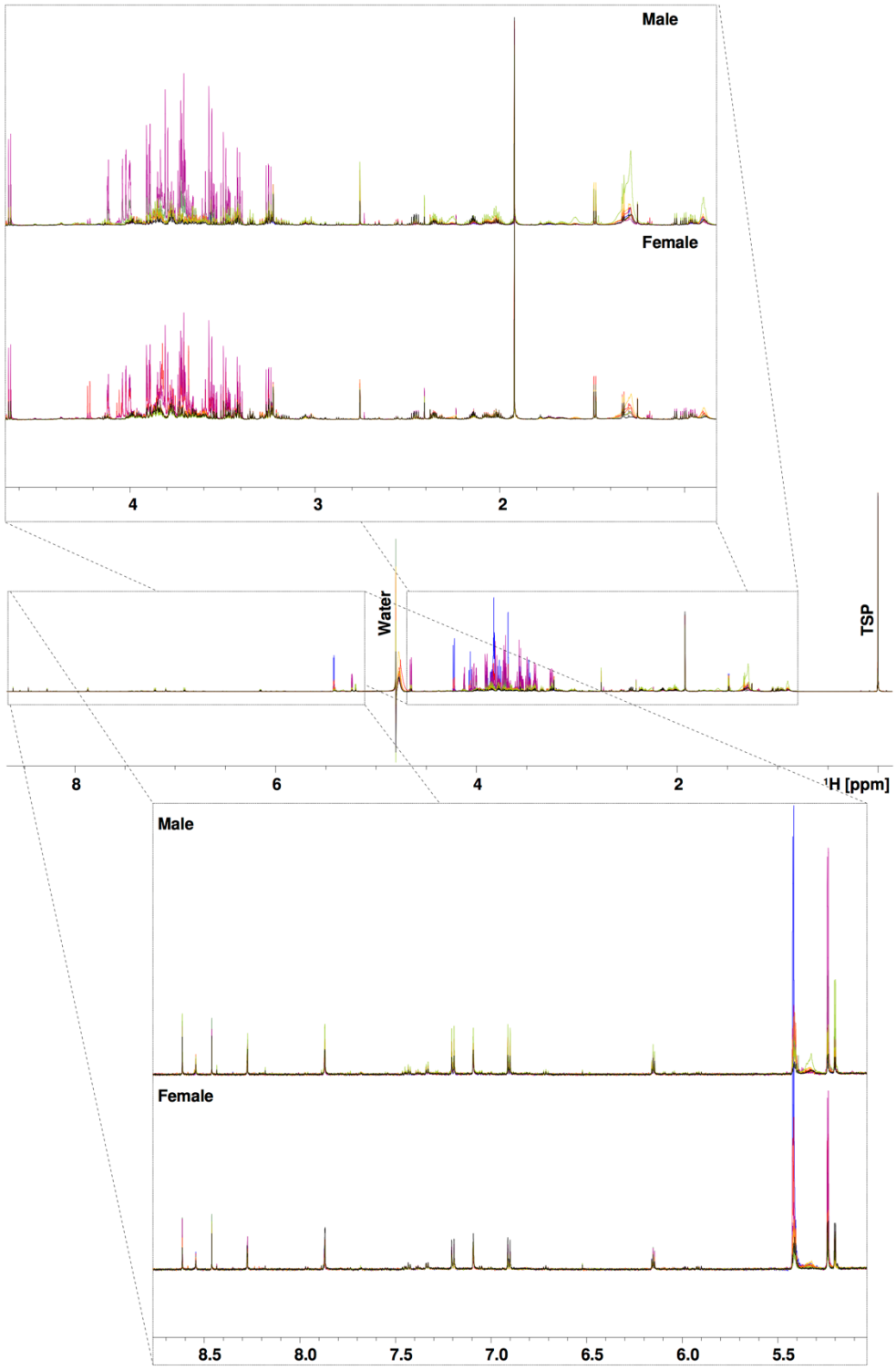
Appendix 3: Metabolite assignment on a ^{1}H-^{13}C HSQC spectrum from a representative *An. gambiae* wild type sample.



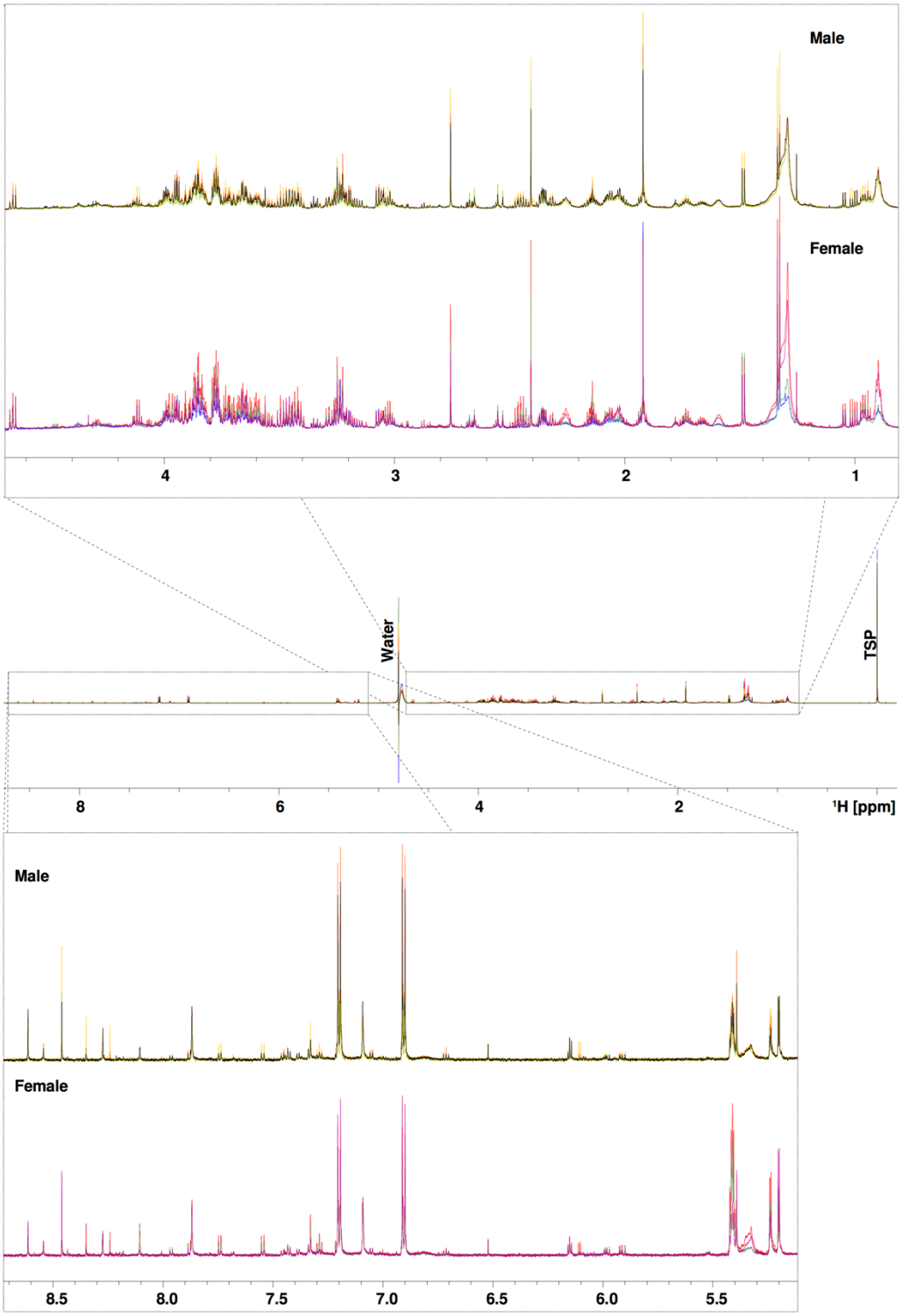
Appendix 4: Metabolite assignment on a ^{1}H-^{13}C HSQC spectrum from a representative *Ae. aegypti* wild type sample.



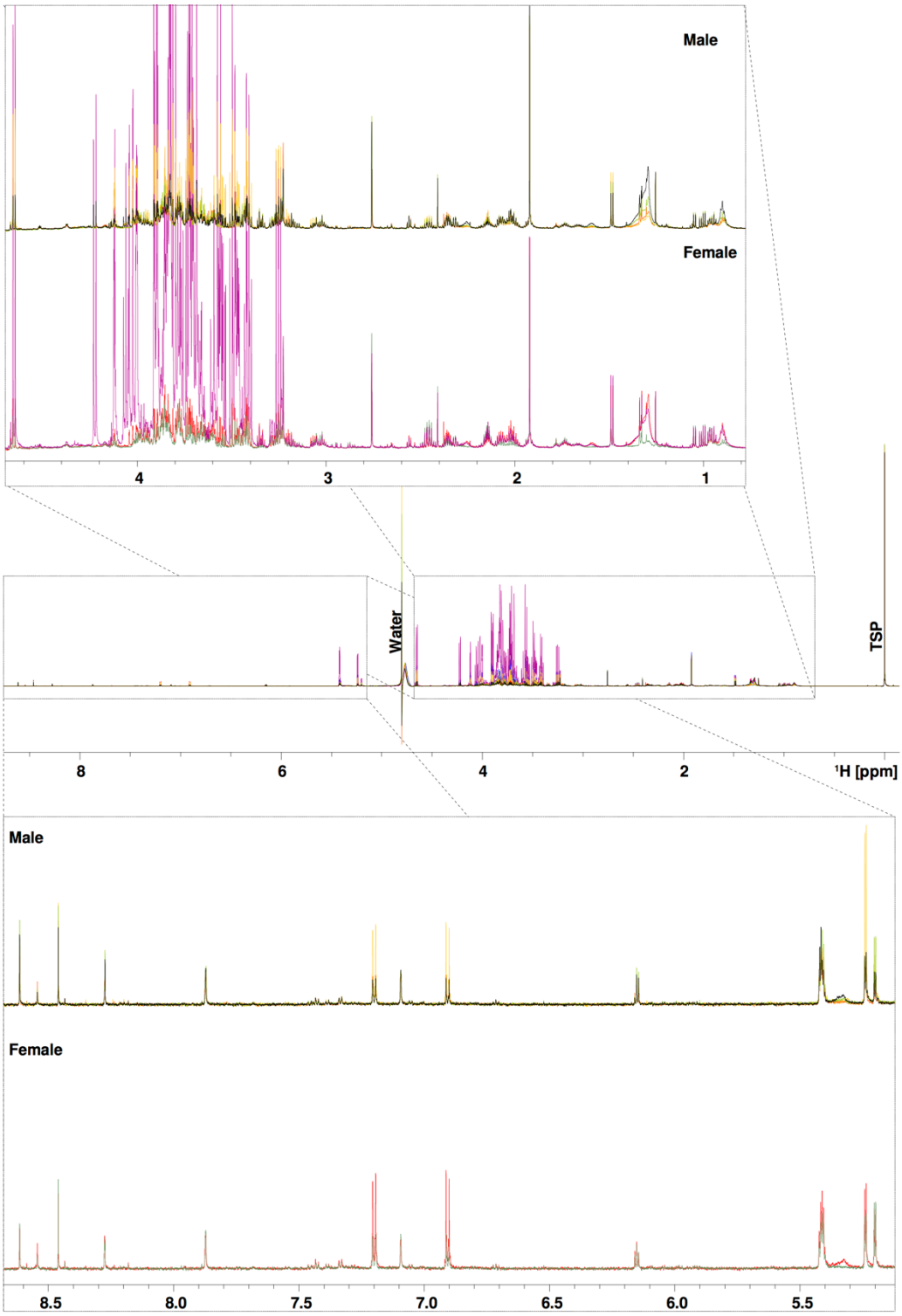
Appendix 5: Representative raw ^{1}H NMR spectra for *An. gambiae* knock-down pupa grouped by sex.



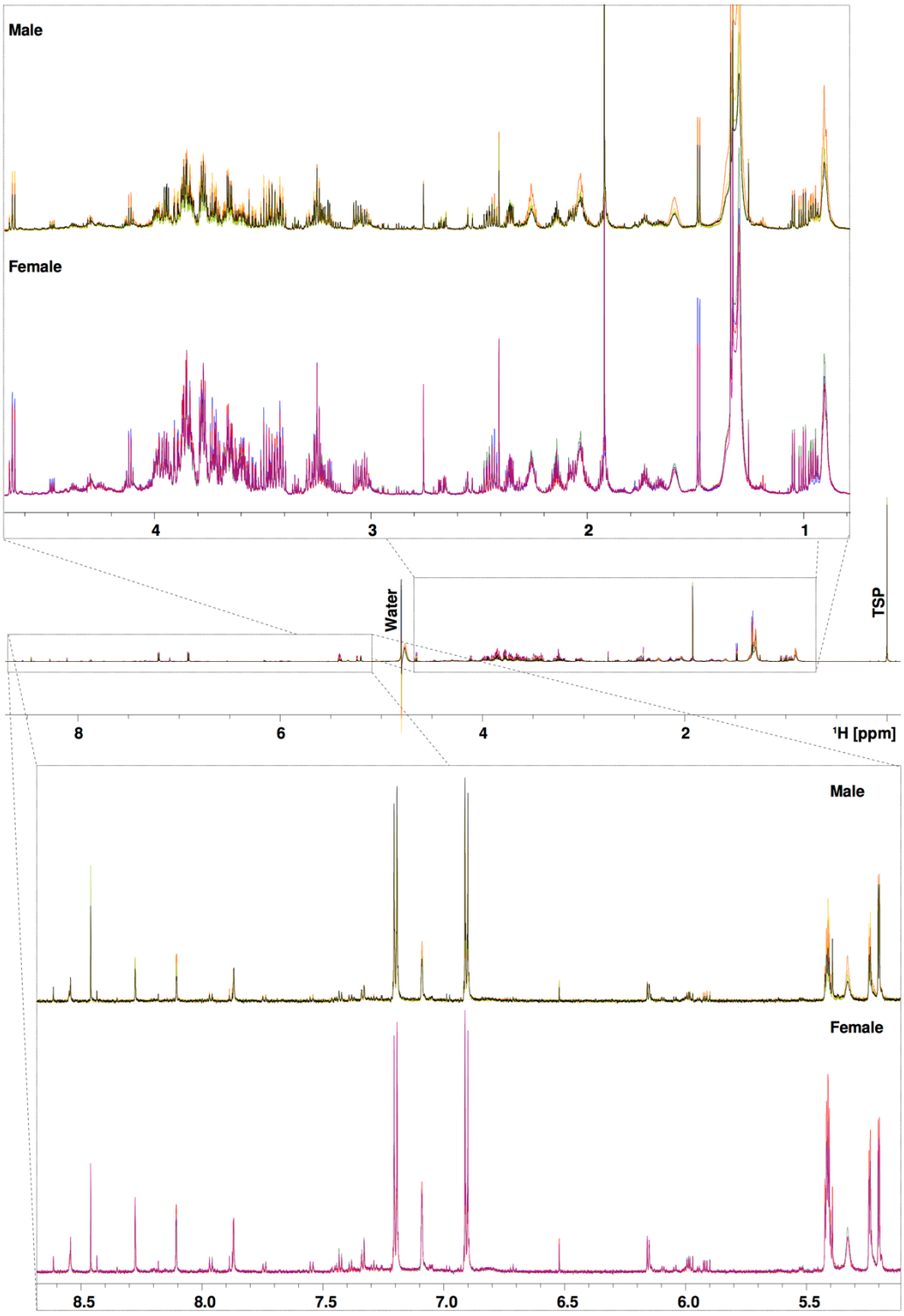
Appendix 6: Representative raw ^{1}H NMR spectra for *An. gambiae* knock-down adult grouped by sex.



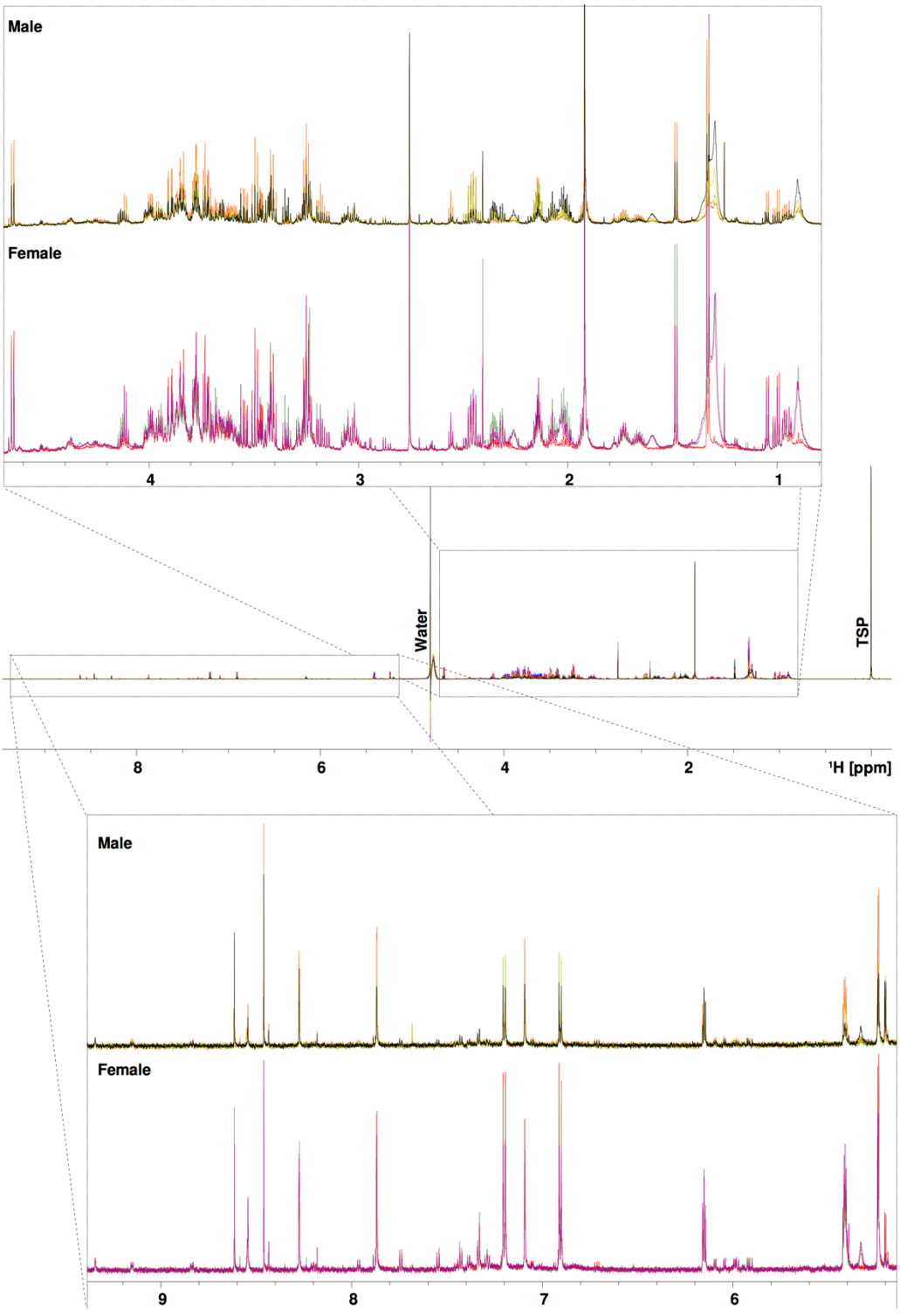
Appendix 7: Representative raw ^{1}H NMR spectra for *An. gambiae* wild type pupa grouped by sex.



Appendix 8: Representative raw ^{1}H NMR spectra for *An. gambiae* wild type adult grouped by sex.



Appendix 9: Representative raw ^{1}H NMR spectra for *Ae. aegypti* wild type pupa grouped by sex.



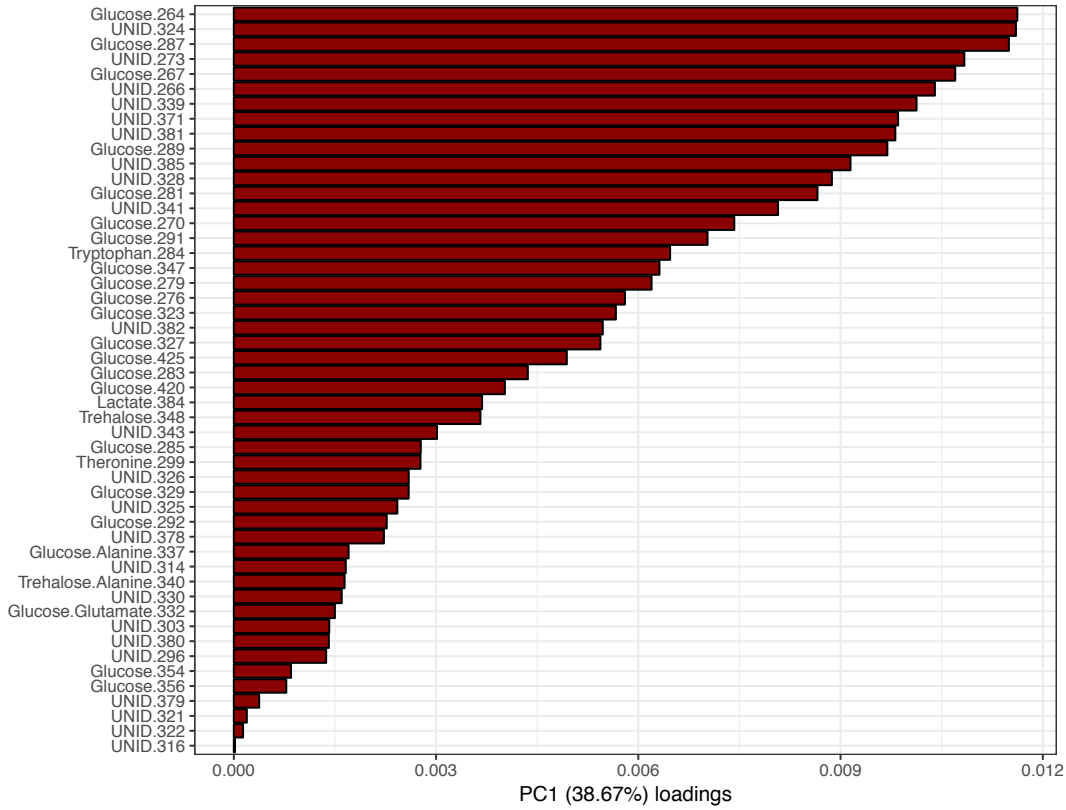
Appendix 10: Representative raw ^{1}H NMR spectra for *Ae. aegypti* wild type adult grouped by sex.

Appendix 11: PLS-DA model assessment metrics. A: adult; Acc: accuracy; Aed: *Aedes Aegypti*; Aga: *Anopheles gambiae*; Av: average; Gen: genotype; KD: knock-down; NA: not applicable; P: pupa; Pre: precision; Rec: recall; Res: resistance; Sel: selected; Spe: species; Sta: stage WT: wild type.

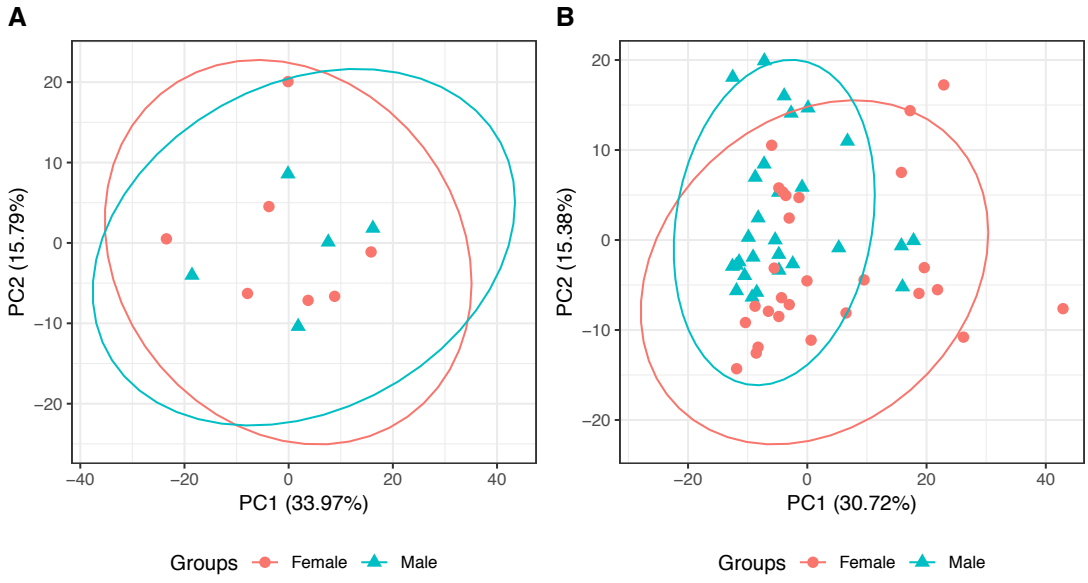
Spe	Gen	Sta	Model	Data	Class	Acc	Pre	Rec	Av Acc	Av pre	Av rec
Aga	KD	P	Sex	Full	NA	0.4545	0.3333	0.2000	NA	NA	NA
Aga	KD	A	Sex	Full	NA	0.9412	0.9500	0.9500	NA	NA	NA
Aga	WT	P	Sex	Full	NA	0.7143	0.7500	0.5000	NA	NA	NA
Aga	WT	A	Sex	Full	NA	0.8750	0.8571	0.8571	NA	NA	NA
Aed	WT	P	Sex	Full	NA	0.9375	1.0000	0.8333	NA	NA	NA
Aed	WT	A	Sex	Full	NA	1.0000	1.0000	1.0000	NA	NA	NA
Aga	KD	P	3-way	Full	KD16	0.9091	0.7500	1.0000	0.7576	0.7500	0.6667
Aga	KD	P	3-way	Full	KD17	0.6364	0.5000	0.7500	0.7576	0.7500	0.6667
Aga	KD	P	3-way	Full	CONT	0.7273	1.0000	0.2500	0.7576	0.7500	0.6667
Aga	KD	A	3-way	Full	KD16	0.7941	0.8333	0.4545	0.8235	0.7341	0.7126
Aga	KD	A	3-way	Full	KD17	0.7941	0.5455	0.7500	0.8235	0.7341	0.7126
Aga	KD	A	3-way	Full	CONT	0.8824	0.8235	0.9333	0.8235	0.7341	0.7126
Aga	KD	P	2-way	Full	NA	0.7500	1.0000	1.0000	NA	NA	NA
Aga	KD	A	2-way	Full	NA	1.0000	1.0000	1.0000	NA	NA	NA
Aga	WT	P	Res	Full	NA	0.6429	1.0000	0.5833	NA	NA	NA
Aga	WT	A	Res	Full	NA	0.8750	0.9000	0.9000	NA	NA	NA
Aed	WT	P	Res	Full	NA	1.0000	1.0000	1.0000	NA	NA	NA
Aed	WT	A	Res	Full	NA	1.0000	1.0000	1.0000	NA	NA	NA
Aga	KD	P	Sex	Sel	NA	0.6364	0.6667	0.6667	NA	NA	NA
Aga	KD	A	Sex	Sel	NA	0.9412	1.0000	0.8824	NA	NA	NA
Aga	WT	P	Sex	Sel	NA	0.7143	0.7273	0.8889	NA	NA	NA
Aga	WT	A	Sex	Sel	NA	0.9375	0.8333	0.9091	NA	NA	NA
Aed	WT	P	Sex	Sel	NA	1.0000	1.0000	1.0000	NA	NA	NA
Aed	WT	A	Sex	Sel	NA	1.0000	1.0000	1.0000	NA	NA	NA
Aga	KD	P	3-way	Sel	KD16	0.9091	1.0000	0.6667	0.6970	0.6667	0.5778
Aga	KD	P	3-way	Sel	KD17	0.5455	0.3333	0.6667	0.6970	0.6667	0.5778
Aga	KD	P	3-way	Sel	CONT	0.6364	0.6667	0.4000	0.6970	0.6667	0.5778
Aga	KD	A	3-way	Sel	KD16	0.7941	0.8000	0.4000	0.7647	0.6686	0.6543
Aga	KD	A	3-way	Sel	KD17	0.7941	0.5000	0.8571	0.7647	0.6686	0.6543
Aga	KD	A	3-way	Sel	CONT	0.7059	0.7059	0.7059	0.7647	0.6686	0.6543
Aga	KD	P	2-way	Sel	NA	0.7500	1.0000	0.3333	NA	NA	NA
Aga	KD	A	2-way	Sel	NA	0.9000	1.0000	0.9412	NA	NA	NA
Aga	WT	P	Res	Sel	NA	0.8571	0.7500	0.7500	NA	NA	NA
Aga	WT	A	Res	Sel	NA	0.9375	0.9000	1.0000	NA	NA	NA
Aed	WT	P	Res	Sel	NA	0.8750	0.8000	1.0000	NA	NA	NA
Aed	WT	A	Res	Sel	NA	0.9167	0.8333	1.0000	NA	NA	NA

Appendix 12: t-Test results for sex comparison. AnKD: *Anopheles gambiae* knock-down, AnWT: *Anopheles gambiae* wild type, AeWT: *Aedes Aegypti* Wild type, DF: degrees of freedom, Raw pval: Raw p-value, BH pval: Benjamini Hochberg adjusted p-value, *: overlapping bin.

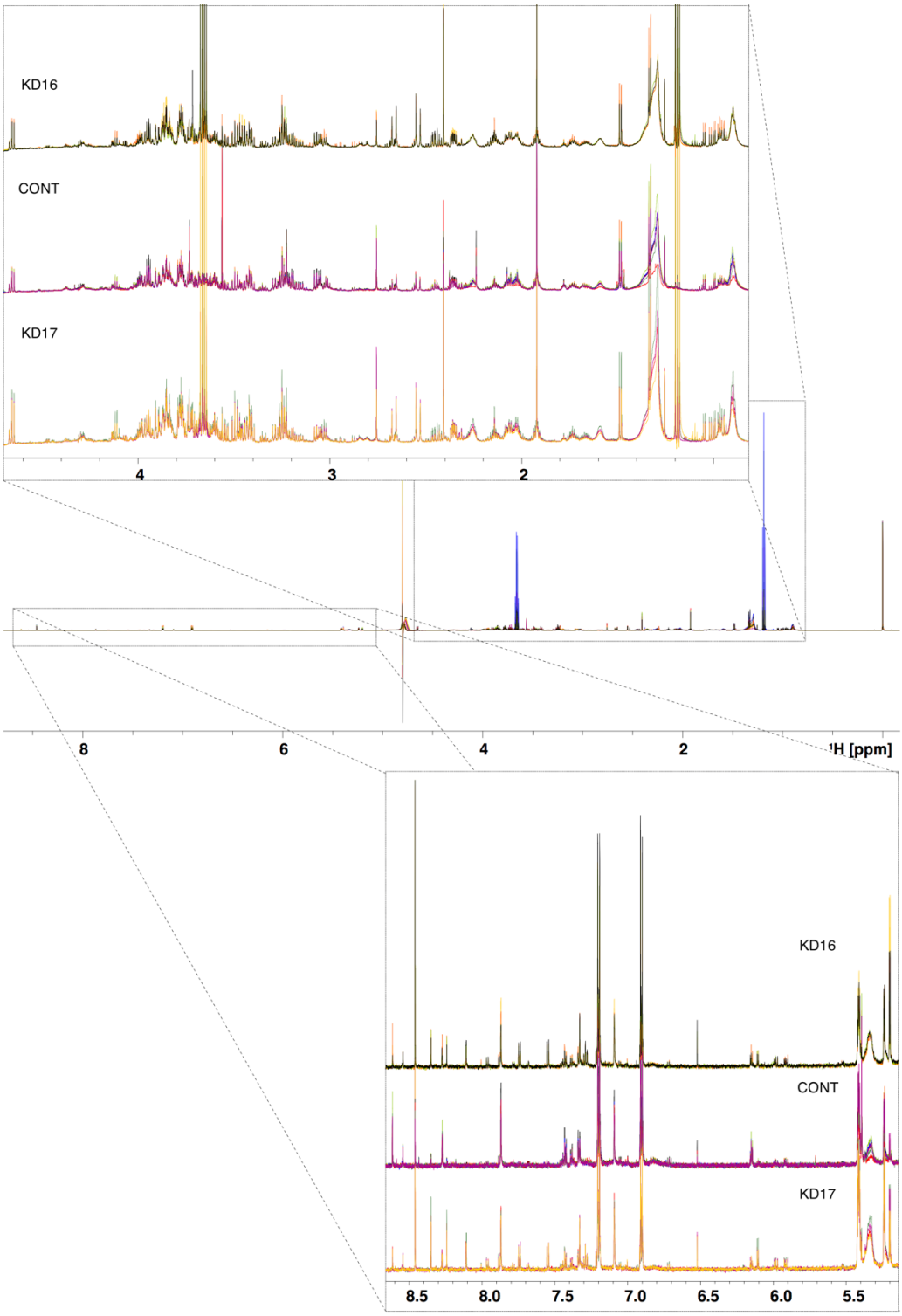
Strain	Stage	Metabolite bin	DF	t-score	Raw pval	BH pval	Female mean	Male mean
AnKD	Pupa	Propionate [17]	34.79	-1.30	2.01E-01	3.61E-01	7.30E+07	7.84E+07
AnKD	Pupa	Acetate [102]	33.91	-2.31	2.70E-02	2.43E-01	2.02E+09	2.22E+09
AnKD	Pupa	Tyrosine [214]	35.90	-0.37	7.11E-01	8.93E-01	2.45E+08	2.55E+08
AnKD	Pupa	Methanol [260]	34.29	1.52	1.39E-01	3.12E-01	7.26E+07	6.51E+07
AnKD	Pupa	Glucose [267]	34.66	0.14	8.93E-01	8.93E-01	5.08E+08	5.03E+08
AnKD	Pupa	Glycine [295]	24.02	-1.82	8.15E-02	3.12E-01	1.61E+08	1.72E+08
AnKD	Pupa	Tryptophan [476]	29.86	-1.14	2.62E-01	3.93E-01	2.98E+07	3.84E+07
AnKD	Pupa	Oxypurinol [493]	33.23	-0.21	8.36E-01	8.93E-01	1.02E+08	1.06E+08
AnKD	Pupa	Formate [497]	26.68	-1.68	1.04E-01	3.12E-01	1.14E+08	1.63E+08
AnKD	Adult	Propionate [12]	92.89	-4.18	6.65E-05	7.98E-05	8.62E+07	9.71E+07
AnKD	Adult	Lactate [386]	107.69	12.37	2.09E-22	1.25E-21	1.85E+08	9.12E+07
AnKD	Adult	Trehalose [424]	78.35	-3.34	1.30E-03	1.30E-03	1.52E+08	2.34E+08
AnKD	Adult	Tyrosine [452]	109.46	-4.95	2.76E-06	4.14E-06	6.61E+07	8.41E+07
AnKD	Adult	Xanthine [482]	107.40	5.20	9.58E-07	1.92E-06	1.96E+07	1.61E+07
AnKD	Adult	Oxypurinol [493]	89.84	-5.46	4.11E-07	1.23E-06	1.40E+08	1.81E+08
AnWT	Pupa	Acetate [102]	44.82	-3.50	1.06E-03	1.82E-03	1.84E+09	2.15E+09
AnWT	Pupa	Propionate [129]	41.81	-4.47	5.89E-05	2.35E-04	7.32E+07	8.38E+07
AnWT	Pupa	Glutamate [138]	45.00	-3.52	1.01E-03	1.82E-03	3.14E+08	3.60E+08
AnWT	Pupa	Pyruvate [139]	44.58	-2.41	1.99E-02	2.17E-02	2.40E+08	2.75E+08
AnWT	Pupa	Glucose [270]	44.68	5.28	3.71E-06	4.00E-05	4.69E+08	3.86E+08
AnWT	Pupa	Valine [304]	44.22	4.16	1.43E-04	3.43E-04	3.08E+08	2.62E+08
AnWT	Pupa	Glutamine [333]*	43.76	2.73	9.09E-03	1.09E-02	1.13E+09	1.03E+09
AnWT	Pupa	Alanine [342]	41.81	5.15	6.66E-06	4.00E-05	2.16E+08	1.93E+08
AnWT	Pupa	Trehalose [346]*	44.86	3.26	2.11E-03	2.82E-03	7.89E+08	7.15E+08
AnWT	Pupa	Lactate [384]	43.92	4.19	1.33E-04	3.43E-04	2.04E+08	1.66E+08
AnWT	Pupa	Fumarate [450]	43.35	-2.17	3.57E-02	3.57E-02	2.74E+07	3.14E+07
AnWT	Pupa	Formate [497]	42.88	-3.42	1.39E-03	2.09E-03	1.10E+08	1.54E+08
AnWT	Adult	Propionate [12]	48.66	-3.71	5.40E-04	1.08E-03	8.26E+07	9.65E+07
AnWT	Adult	Acetate [102]	44.09	-4.95	1.14E-05	3.04E-05	1.92E+09	2.25E+09
AnWT	Adult	Glycine [295]	27.47	3.19	3.57E-03	4.38E-03	1.17E+09	3.86E+08
AnWT	Adult	Alanine [342]	28.40	3.46	1.71E-03	2.74E-03	6.01E+08	2.58E+08
AnWT	Adult	Trehalose [351]	32.92	3.07	4.29E-03	4.38E-03	8.89E+08	6.66E+08
AnWT	Adult	Glucose [354]	30.10	3.08	4.38E-03	4.38E-03	2.93E+09	1.19E+09
AnWT	Adult	Lactate [386]	38.83	9.02	4.54E-11	3.63E-10	2.55E+08	1.08E+08
AnWT	Adult	Oxypurinol [493]	40.60	-5.80	8.65E-07	3.46E-06	1.51E+08	2.27E+08
AeWT	Pupa	Isoleucine [4]	50.25	-3.47	1.09E-03	1.09E-03	5.35E+08	5.85E+08
AeWT	Pupa	Acetate [114]	46.36	-9.47	2.09E-12	1.05E-11	3.60E+09	5.63E+09
AeWT	Pupa	Propionate [141]	47.92	-10.95	1.23E-14	1.23E-13	6.88E+07	8.66E+07
AeWT	Pupa	Glucose [284]	49.39	4.27	8.84E-05	1.10E-04	6.72E+08	5.18E+08
AeWT	Pupa	Glycine [309]	47.48	6.11	1.75E-07	4.38E-07	2.96E+08	2.23E+08
AeWT	Pupa	Alanine [356]	52.40	4.69	1.97E-05	2.82E-05	2.85E+08	2.50E+08
AeWT	Pupa	Trehalose [362]	50.15	3.83	3.59E-04	3.99E-04	1.74E+09	1.50E+09
AeWT	Pupa	Lactate [398]	48.62	4.85	1.30E-05	2.53E-05	3.01E+08	2.57E+08
AeWT	Pupa	Tyrosine [466]	52.55	-4.77	1.52E-05	2.53E-05	4.04E+08	5.15E+08
AeWT	Pupa	Formate [514]	41.20	-9.40	8.24E-12	2.75E-11	7.46E+07	1.34E+08
AeWT	Adult	Valine [8]	34.68	8.56	4.55E-10	1.25E-09	3.90E+08	2.35E+08
AeWT	Adult	Isoleucine [9]	31.37	5.30	8.67E-06	1.59E-05	1.83E+08	1.37E+08
AeWT	Adult	Propionate [15]	24.83	-13.95	2.98E-13	3.28E-12	1.33E+08	2.40E+08
AeWT	Adult	Acetate [114]	23.66	-12.20	1.08E-11	5.92E-11	5.74E+09	1.02E+10
AeWT	Adult	Pyruvate [151]	25.44	-3.22	3.50E-03	3.85E-03	1.38E+08	1.95E+08
AeWT	Adult	Methanol [274]	25.35	-3.11	4.61E-03	4.61E-03	1.32E+08	2.06E+08
AeWT	Adult	Lactate [398]	38.71	5.64	1.65E-06	3.63E-06	2.56E+08	2.05E+08
AeWT	Adult	Glucose [434]	38.16	3.86	4.30E-04	5.92E-04	3.77E+08	2.94E+08
AeWT	Adult	Trehalose [438]	25.96	-3.73	9.55E-04	1.17E-03	1.24E+08	2.66E+08
AeWT	Adult	Tryptophan [483]	35.76	3.91	3.98E-04	5.92E-04	1.33E+08	1.14E+08
AeWT	Adult	Formate [514]	23.56	-11.93	1.81E-11	6.63E-11	1.27E+08	2.46E+08



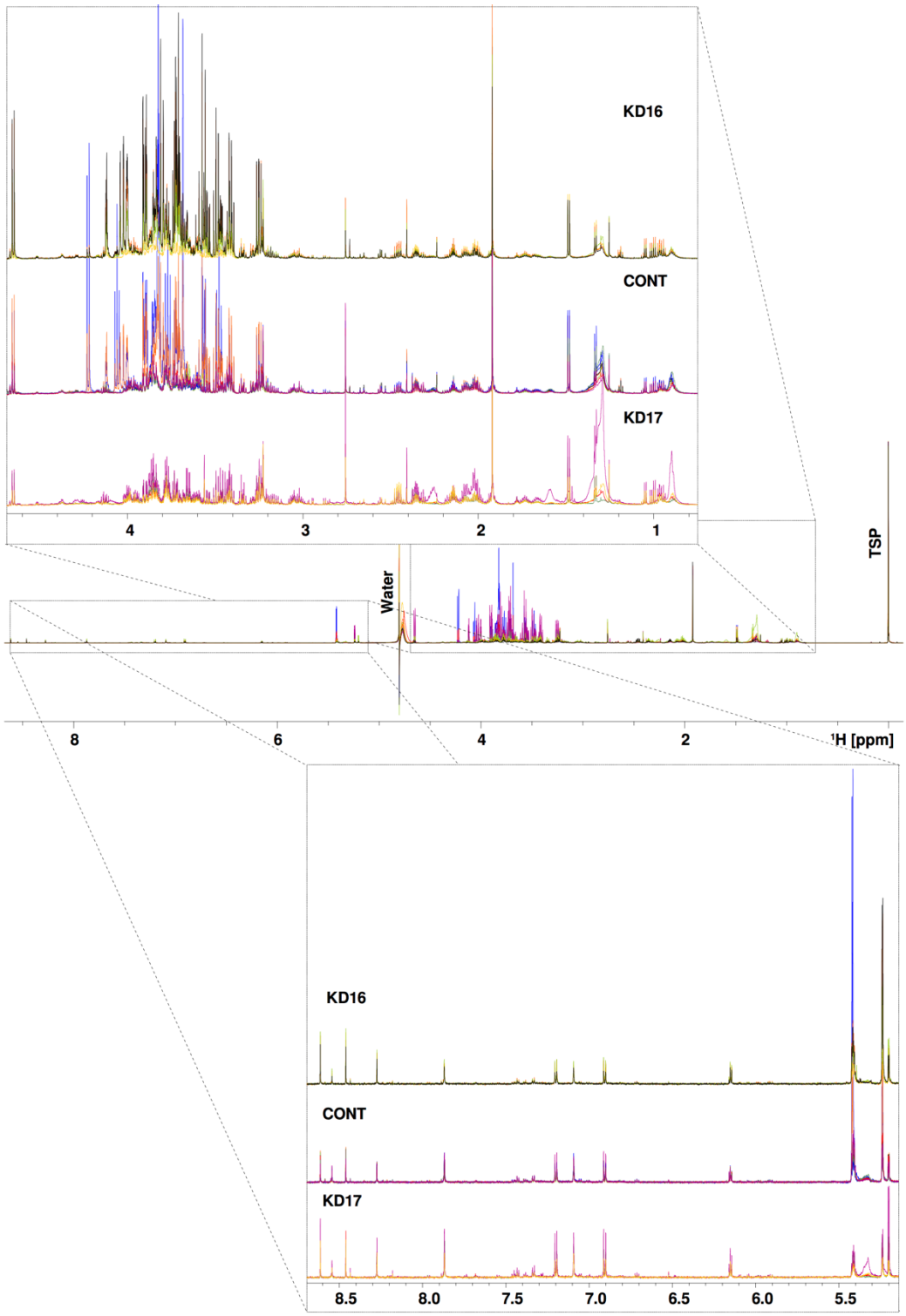
Appendix 13: Wild type An. gambiae loadings for PC1. The sub-population observed in the PCA scores plot is explained by PC1. The most influential (first 50) bins for PC1 is explained by 22 glucose bins. UNID: unidentified bin.



Appendix 14: *An. gambiae* knock-down control samples PCA scores (A: pupae, B: adults) highlighted by sex. Both pupa and adult controls does not show a clear separation of sex as observed in wild type *An. gambiae*. Ellipses represent 95% confidence region.



Appendix 15: Representative raw ^{1}H NMR spectra for *An. gambiae* knock-down pupa grouped by genotype. KD16: Cyp4g16 knock-down, KD17: Cyp4g17 knock-down, CONT: Control strain.



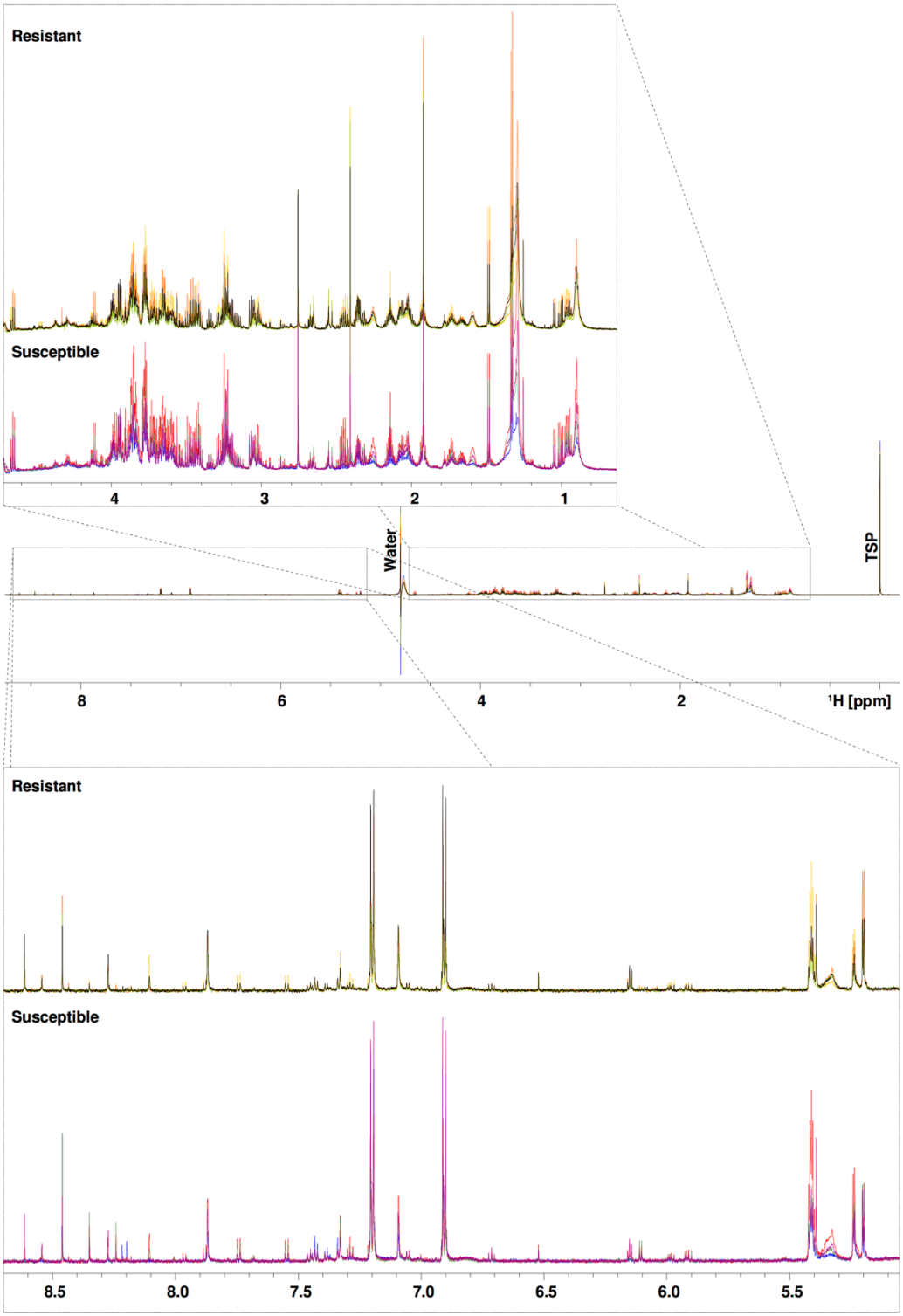
Appendix 16: Representative raw ^{1}H NMR spectra for *An. gambiae* knock-down adult grouped by genotype. KD16: Cyp4g16 knock-down, KD17: Cyp4g17 knock-down, CONT: Control strain.

Appendix 17: ANOVA, Tukey's HSD and t-test results for knock-down comparisons. DF: degrees of freedom, Raw pval: Raw p-value, BH pval: Benjamini Hochberg adjusted p-value, S: stage, *: overlapping bin.

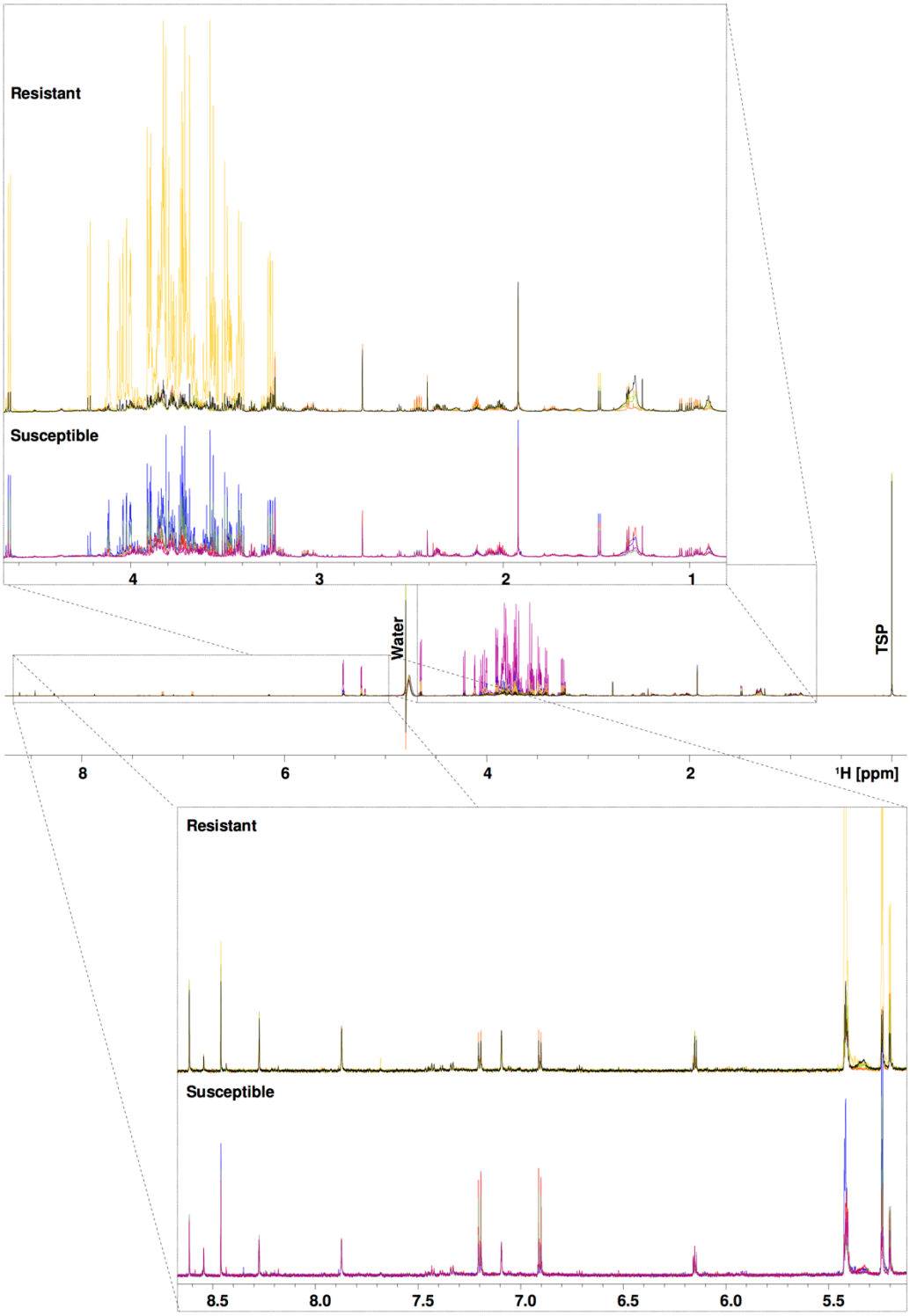
		ANOVA				Tukey's HSD		
S	Metabolite bin	DF	F-value	Raw pval	BH pval	KD16: CONT	KD17: CONT	KD17: KD16
P	Valine [8]	2	8.74	8.34E-04	2.92E-03	5.35E-04	1.29E-01	4.98E-02
P	Propionate [17]	2	4.81	1.43E-02	2.50E-02	2.66E-02	9.91E-01	2.56E-02
P	Tyrosine [207]	2	0.78	4.65E-01	4.65E-01	5.41E-01	5.14E-01	9.99E-01
P	Alanine & Glutamate [334]	2	3.00	6.26E-02	8.76E-02	2.85E-01	5.23E-02	7.36E-01
P	Trehalose [351]	2	5.91	6.17E-03	1.44E-02	5.69E-03	5.90E-01	4.06E-02
P	Glucose [425]	2	2.55	9.25E-02	1.08E-01	2.28E-01	9.24E-02	9.37E-01
P	Tryptophan [476]	2	10.07	3.52E-04	2.47E-03	5.63E-04	2.85E-03	6.59E-01
A	Valine [10]	2	12.00	1.95E-05	3.25E-05	5.71E-03	3.66E-05	4.87E-01
A	Tyrosine [207]	2	17.04	3.64E-07	1.82E-06	5.91E-01	2.44E-07	2.69E-04
A	Trehalose [311]	2	15.33	1.36E-06	3.40E-06	3.48E-02	6.75E-04	8.77E-07
A	Glucose [329]	2	11.19	3.81E-05	4.10E-05	1.13E-01	3.04E-03	2.85E-05
A	Alanine [340]*	2	11.10	4.10E-05	4.10E-05	6.31E-02	6.31E-03	2.53E-05

t-test

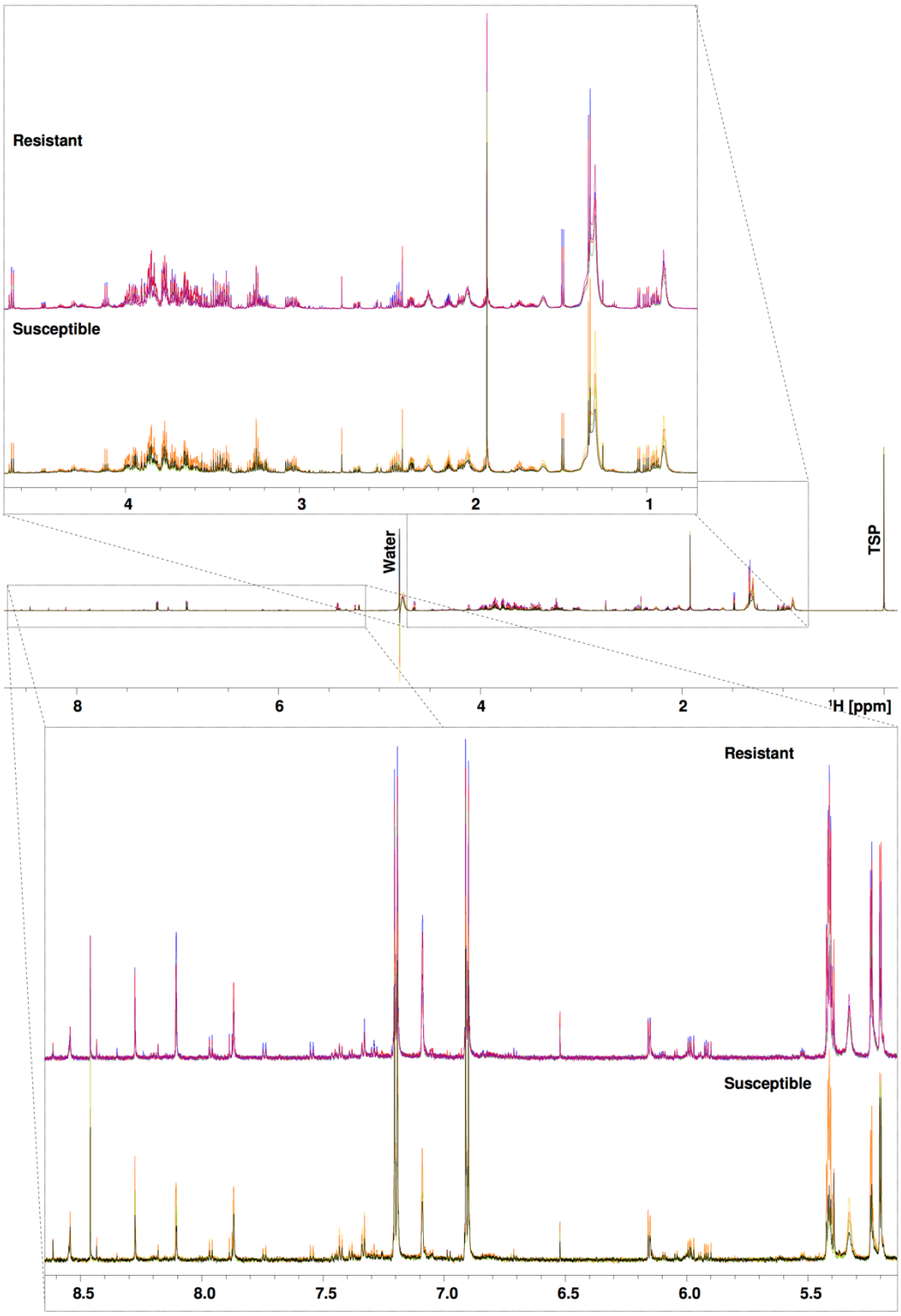
S	Metabolite bin	DF	t-score	Raw pval	BH pval	KD16 mean	KD17 mean
P	Valine [13]	19.48	1.99	6.07E-02	1.45E-01	3.15E+08	2.68E+08
P	Propionate [17]	23.68	2.68	1.30E-02	1.12E-01	8.47E+07	7.17E+07
P	Glutamine [126]*	12.89	2.43	3.06E-02	1.12E-01	2.31E+08	1.99E+08
P	Pyruvate [139]	16.37	-0.74	4.70E-01	5.25E-01	2.03E+08	2.16E+08
P	Glutamate [331]	15.06	1.32	2.08E-01	3.81E-01	3.93E+08	3.59E+08
P	Alanine [334]*	22.25	-0.65	5.25E-01	5.25E-01	7.03E+08	7.30E+08
P	Trehalose [351]	15.19	1.98	6.61E-02	1.45E-01	6.14E+08	5.21E+08
P	Glucose [354]	23.81	0.99	3.33E-01	4.58E-01	5.36E+08	5.01E+08
P	Tryptophan [476]	20.53	0.72	4.81E-01	5.25E-01	4.61E+07	3.97E+07
P	Xanthine [482]	19.96	-2.49	2.17E-02	1.12E-01	2.37E+07	2.77E+07
P	Formate [497]	16.33	1.10	2.88E-01	4.53E-01	1.89E+08	1.49E+08
A	Propionate [15]	48.05	-3.70	5.59E-04	5.59E-04	9.47E+07	1.26E+08
A	Acetate [102]	53.66	-6.55	2.26E-08	1.81E-07	2.58E+09	3.77E+09
A	Succinate [144]	53.70	4.25	8.52E-05	1.70E-04	2.42E+08	1.73E+08
A	Tyrosine [212]	53.43	-4.14	1.23E-04	1.97E-04	8.44E+07	1.00E+08
A	Glucose [285]	31.42	4.18	2.18E-04	2.91E-04	7.35E+08	3.13E+08
A	Glycine [295]	35.72	6.51	1.49E-07	5.97E-07	5.10E+08	1.27E+08
A	Trehalose [313]	37.10	3.89	4.01E-04	4.58E-04	8.39E+08	5.18E+08
A	Alanine [342]	34.84	4.62	5.13E-05	1.37E-04	3.02E+08	1.54E+08



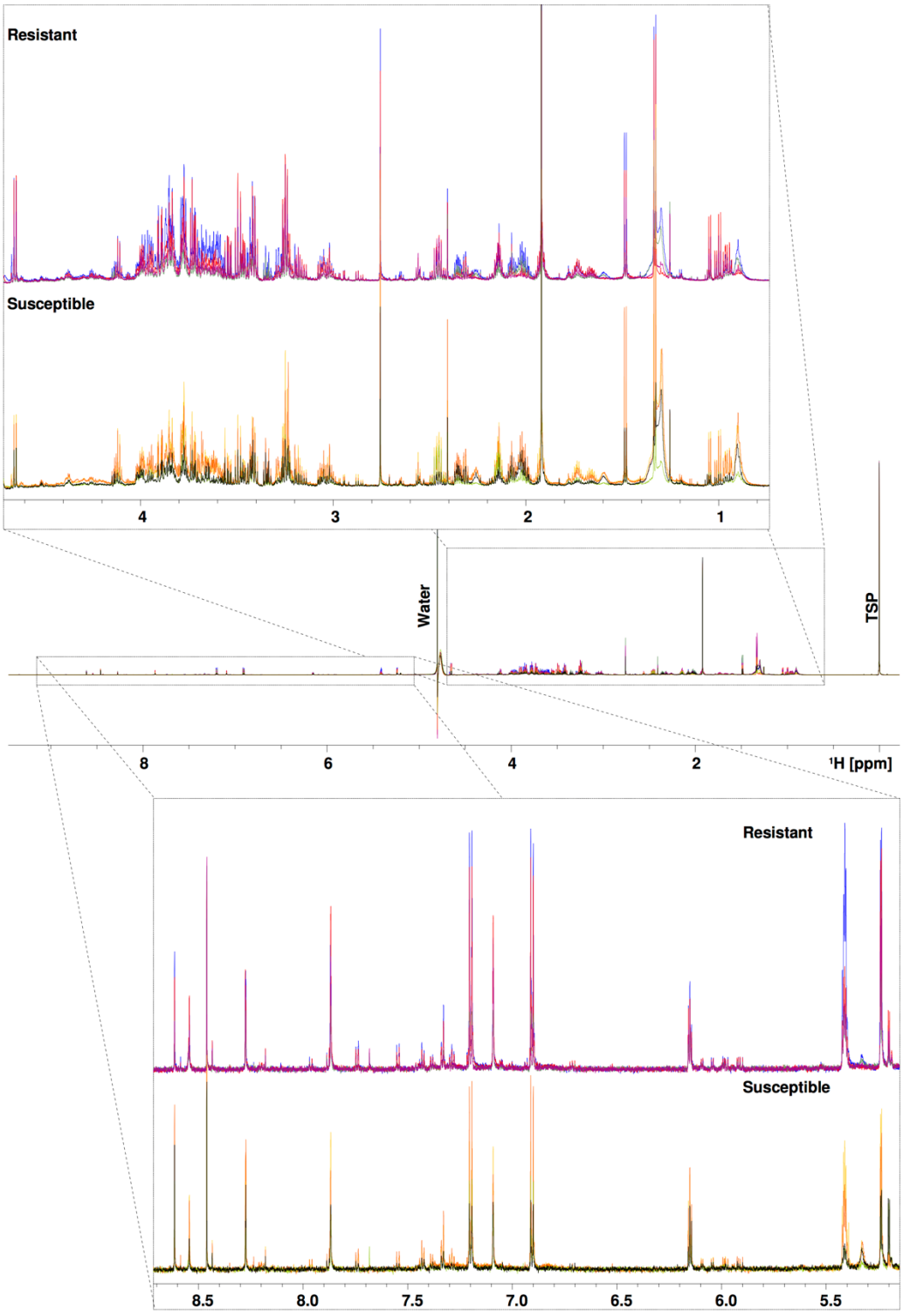
Appendix 18: Representative raw ^{1}H NMR spectra for *An. gambiae* wild type pupa grouped by resistance status.



Appendix 19: Representative raw ^{1}H NMR spectra for *An. gambiae* wild type adult grouped by resistance status.



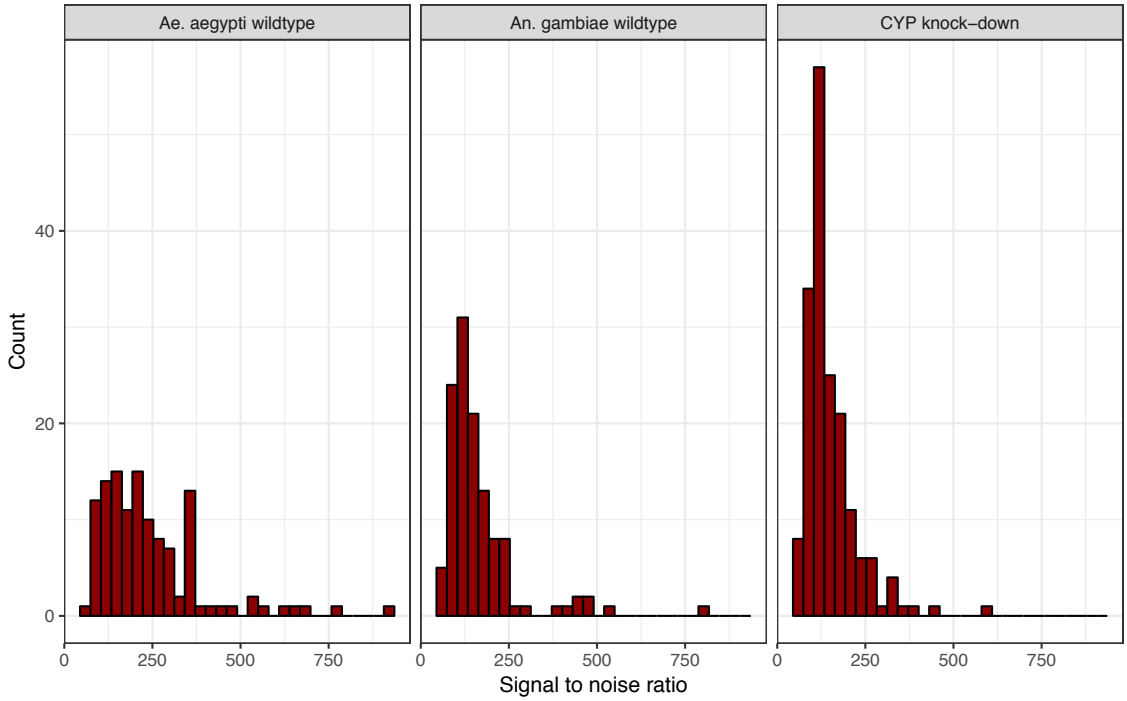
Appendix 20: Representative raw ^{1}H NMR spectra for *Ae. aegypti* wild type pupa grouped by resistance status.



Appendix 21: Representative raw ^{1}H NMR spectra for *Ae. aegypti* wild type adult grouped by resistance status.

Appendix 22: Wild type *Anopheles gambiae* and *Aedes aegypti* resistant v susceptible t-test results. Ano: *Anopheles gambiae*, Aed: *Aedes Aegypti*, DF: degrees of freedom, Raw pval: Raw p-value, BH pval: Benjamini Hochberg adjusted p-value, Res: resistant, Sus: susceptible, *: overlapping bin.

Species	Stage	Metabolite bin	DF	t-score	Raw pval	BH pval	Res mean	Sus mean
Ano	Pupa	Isoleucine [4]	41.33	-3.21	2.53E-03	2.89E-03	4.48E+08	4.94E+08
Ano	Pupa	Acetate [102]	36.26	2.85	7.24E-03	7.73E-03	2.15E+09	1.88E+09
Ano	Pupa	Propionate [129]	42.26	2.71	9.64E-03	9.64E-03	8.26E+07	7.53E+07
Ano	Pupa	Pyruvate [139]	34.21	9.70	2.37E-11	3.79E-10	3.06E+08	2.21E+08
Ano	Pupa	Tryptophan [252]	44.58	3.48	1.15E-03	1.67E-03	9.98E+07	8.06E+07
Ano	Pupa	Methanol [260]	33.23	8.39	1.02E-09	8.17E-09	9.92E+07	6.86E+07
Ano	Pupa	Glucose [270]	35.65	-4.58	5.54E-05	1.48E-04	3.84E+08	4.62E+08
Ano	Pupa	Trehalose [275]	25.23	3.57	1.46E-03	1.95E-03	3.59E+08	2.98E+08
Ano	Pupa	Glycine [295]	39.54	-3.64	7.84E-04	1.25E-03	1.68E+08	1.91E+08
Ano	Pupa	Threonine [299]	38.40	-4.55	5.28E-05	1.48E-04	3.53E+08	4.34E+08
Ano	Pupa	Glutamate [331]	28.67	4.04	3.61E-04	7.22E-04	4.68E+08	4.18E+08
Ano	Pupa	Alanine [342]	41.11	-4.32	9.76E-05	2.23E-04	1.93E+08	2.14E+08
Ano	Pupa	Lactate [384]	36.39	-4.99	1.53E-05	6.12E-05	1.60E+08	2.05E+08
Ano	Pupa	Fumarate [450]	42.16	7.19	7.59E-09	4.05E-08	3.49E+07	2.53E+07
Ano	Pupa	Oxypurinol [493]	43.99	3.66	6.69E-04	1.19E-03	1.39E+08	9.07E+07
Ano	Pupa	Formate [497]	34.52	3.31	2.21E-03	2.72E-03	1.57E+08	1.13E+08
Ano	Adult	Lactate [49]	49.71	-2.70	9.52E-03	1.07E-02	8.95E+08	1.17E+09
Ano	Adult	Propionate [129]	45.86	2.82	7.01E-03	9.19E-03	5.97E+07	5.35E+07
Ano	Adult	Pyruvate [139]	49.45	-5.57	1.06E-06	9.50E-06	1.38E+08	1.80E+08
Ano	Adult	Succinate [144]	49.68	3.75	4.59E-04	1.03E-03	2.77E+08	2.29E+08
Ano	Adult	Methanol [260]	49.98	-4.37	6.33E-05	1.90E-04	1.42E+08	1.80E+08
Ano	Adult	Trehalose [315]	50.87	2.80	7.15E-03	9.19E-03	8.68E+08	6.21E+08
Ano	Adult	Theronine [398]	50.48	-3.52	9.29E-04	1.67E-03	6.24E+07	6.97E+07
Ano	Adult	Tyrosine [452]	31.58	-2.11	4.28E-02	4.28E-02	8.14E+07	1.03E+08
Ano	Adult	Oxypurinol [493]	50.97	4.64	2.45E-05	1.10E-04	2.21E+08	1.54E+08
Aed	Pupa	Alanine [58]	51.84	8.09	9.40E-11	4.70E-10	7.05E+08	5.63E+08
Aed	Pupa	Pyruvate [151]	45.37	-6.89	1.44E-08	4.66E-08	2.42E+08	3.04E+08
Aed	Pupa	Methanol [274]	48.25	-6.43	5.44E-08	1.09E-07	6.79E+07	8.81E+07
Aed	Pupa	Glucose [284]	46.02	9.25	4.63E-12	4.63E-11	7.07E+08	4.70E+08
Aed	Pupa	Glycine [309]	48.54	6.72	1.87E-08	4.66E-08	2.95E+08	2.19E+08
Aed	Pupa	Trehalose [365]	50.02	4.43	5.22E-05	5.22E-05	1.03E+09	8.99E+08
Aed	Pupa	Tyrosine [374]	52.79	4.41	5.09E-05	5.22E-05	6.56E+08	5.81E+08
Aed	Pupa	Lactate [398]	47.24	5.23	3.74E-06	5.11E-06	3.01E+08	2.54E+08
Aed	Pupa	Xanthine [499]	47.61	6.21	1.25E-07	2.08E-07	4.38E+07	3.44E+07
Aed	Pupa	Oxypurinol [510]	49.62	-5.18	4.09E-06	5.11E-06	6.29E+07	8.19E+07
Aed	Adult	Propionate [139]	38.05	-3.12	3.47E-03	3.82E-03	9.79E+07	1.14E+08
Aed	Adult	Pyruvate [151]	35.28	-4.17	1.86E-04	4.10E-04	1.32E+08	1.99E+08
Aed	Adult	Succinate [156]	33.24	-4.27	1.54E-04	4.10E-04	3.61E+08	4.42E+08
Aed	Adult	Methanol [274]	36.49	-3.93	3.63E-04	6.65E-04	1.24E+08	2.10E+08
Aed	Adult	Valine [318]	37.80	3.16	3.11E-03	3.80E-03	3.78E+08	2.98E+08
Aed	Adult	Alanine [351*]	38.61	7.21	1.19E-08	1.30E-07	1.25E+09	1.00E+09
Aed	Adult	Trehalose [362]	36.58	3.55	1.09E-03	1.71E-03	1.28E+09	1.10E+09
Aed	Adult	Glucose [368]	38.29	4.58	4.84E-05	1.77E-04	9.77E+08	7.97E+08
Aed	Adult	Tyrosine [374]	27.83	3.42	1.94E-03	2.67E-03	4.40E+08	3.80E+08
Aed	Adult	Tryptophan [493]	39.00	2.66	1.12E-02	1.12E-02	3.85E+07	3.26E+07
Aed	Adult	Oxypurinol [510]	31.14	-5.17	1.31E-05	7.20E-05	2.38E+08	3.26E+08



Appendix 23: Signal to noise ratio histogram calculated from all data sets used in this project. Signal region was defined as 1-4 ppm and noise region was defined as 10-10.5 ppm, Region integrations were normalised to the region range prior to Signal to noise calculation.